DEVELOPMENT OF THE UC AUDITORY-VISUAL MATRIX SENTENCE TEST

A thesis submitted in partial fulfilment of the requirements for the Degree

of Master of Audiology

at the University of Canterbury

by Ronald Harris Trounson

University of Canterbury

2012

Acknowledgements

The author wishes to acknowledge his primary supervisor, Dr Greg O'Beirne, for his unwavering support, enthusiasm and vision for the development of an auditory-visual matrix sentence test, something which we believe is the first of its kind in the world. As with any ground breaking research, it would not be possible without the assistance of a truly dedicated team. The author wishes to acknowledge of help of his secondary supervisor, Dr Margaret Maclagan, and linguist Ruth Hope for their expertise and guidance on the subtleties of New Zealand English. The author wishes to acknowledge Professor Nancy Tye-Murray for her inspirational research into auditory-visual enhancement. The author wishes to acknowledge Emma Parnell of the New Zealand Institute of Language, Brain and Behaviour along with Rob Stowell from the University of Canterbury (UC) College of Education and John Chrisstoffels from the UC School of Fine for their video equipment and cinematography expertise. The author wishes to acknowledge the speaker, actress Emma Johnston, for her incredible patience and self control throughout the recording sessions. Last but not least, the author wishes to acknowledge Dr Emily Lin from the UC Department of Communication Disorders for her expertise and guidance on statistical methods.

Abstract

Matrix Sentence Tests consist of syntactically fixed but semantically unpredictable sentences each composed of 5 words (name, verb, quantity, adjective, object). Test sentences are generated by choosing 1 of 10 alternatives for each word to form sentences such as "Amy has nine green shoes". Up to 100,000 unique sentences are possible. Rather than recording these sentences individually, the sentences are synthesized from 400 recorded audio fragments that preserve coarticulations and provide a natural prosody for the synthesized sentence. Originally developed by for the Swedish language in 1982, Matrix Sentence Tests are now available in German, Danish, British English, Polish and Spanish. The Matrix Sentence Test has become the standard speech audiometry measure in much of Europe, and was selected by the HearCom consortium as a means of standardising speech audiometry across the different European regions and languages. Existing Matrix Sentence Tests function in auditory-only mode. We describe the development of a New Zealand English Matrix Sentence Test in which we have made the important step of adding an auditory-visual mode, using recorded fragments of video rather than simply audio. The addition of video stimuli not only increases the face-validity of the test, but allows different presentation modes to be compared, thereby allowing the contribution of visual cues to be assessed.

Table of Contents

List of Figures

List of Tables

Abbreviations

AAE	Adobe After Effects
AM	Amplitude Modulation
AME	Adobe Media Encoder
APP	Adobe Premiere Pro
avi	Audio video interleave file format
BKB-SIN	Bamford-Kowal-Bench Speech-in-Noise Test
CST	Connected Sentence Test
CVC	Consonant-Vowel-Consonant
dB	Decibel
F1	First formant
F2	Second formant
fps	Frames per second
GB	Gigabyte
HD	High Definition
HINT	Hearing-in-Noise Test
Hz	Hertz
IAC	Industrial Acoustics Company Ltd
jpeg	Joint Photographic Experts Group image file format
NZ	New Zealand
NZDTT	New Zealand Digit Triplet Test
NZHINT	New Zealand Hearing-in-Noise Test
m4a	mpeg4 audio format
mp4	mpeg4 video format
mpeg4	Moving Picture Experts Group, standard 4 file format
mpg	Moving Picture Experts Group, standard 1 file format
MB	Megabyte
MS-DOS	Microsoft-Disk Operating System
PAL	Phase Alternating Line
PC	Personal Computer
PCM	Pulse-code Modulation

PI	Performance-Intensity
QuickSIN	Quick Speech-in-Noise
RAM	Random Access Memory
RGB	Red, Green, Blue
SNR	Signal-to-Noise Ratio
SRT	Speech Reception Threshold
UC	University of Canterbury
UCAST	University of Canterbury Adaptive Speech Test
USB	Universal Serial Bus
VI	Virtual Instrument
wav	Waveform audio file format
WIN	Words-in-Noise

1 Introduction

Hearing and understanding speech have unique importance in our lives. For children, the ability to hear and understand speech is fundamental to the development of oral language. For adults, difficulty in detecting and understanding speech limits the ability to participate in the communication interactions that are the foundation of numerous activities of daily living. In 1951 the father of Audiology, Raymond Carhart, defined speech audiometry as "a technique wherein standardized samples of a language are presented through a calibrated system to measure some aspect of hearing ability" (Carhart, 1951). Today, speech audiometry is an integral part of the audiological test battery. It is a key measure of overall auditory perception skills, providing an indication of an individual's ability to identify and discriminate phonetic segments, words, sentences and connected discourse (Mendel, 2008). Scores on speech tests are often used as a crosscheck of the validity of pure-tone thresholds (McArdle & Hnath-Chislom, 2009).

1.1 Speech Audiometry in New Zealand

The materials used in speech audiometry in New Zealand are generally monosyllabic word lists presented in quiet, such as the Meaningful CVC (Consonant-Vowel-Consonant) Words (Boothroyd & Nittrouer, 1988). Items are presented in lists, often after a carrier phrase, such as "say (the word) ____". Words are presented in isolation, without context, so that patients must repeat what they hear without relying on contextual clues. The aim is to attempt to isolate the problem of audibility from other confounding factors such as working memory and use of context (Wilson, McArdle, & Smith, 2007). Performance is scored by word or by phoneme repeated correctly to arrive at a percentage correct score. A number of word lists are presented at two or more different intensity levels in order to describe a performance-intensity (PI) function, from which the speech reception threshold (SRT) or 50% correct point can be estimated. The conditions under which speech audiometry is performed in the clinic are optimal compared to those encountered in the real world. Speech materials are presented

through headphones, in a soundproof room, with maximum concentration from the patient and minimum external distraction.

1.2 Disadvantages of Fixed-Level Monosyllabic Words in Quiet

Speech recognition testing in quiet does not address the main problem experienced by the majority of hearing impaired listeners, which is difficulty understanding speech in noise. Listeners with identical word recognition abilities in quiet can have significantly different word recognition abilities in background noise (Beattie, Barr, & Roup, 1997). The assessment of receptive communication abilities ideally should involve speech materials and listening conditions that are likely to be encountered in the real world.

Speech tests in quiet of the kind that are currently used in audiological practice in New Zealand fall into the category of non-adaptive tests. These methods are susceptible to floor and ceiling effects (where a number of participants obtain scores of, or close to, 0% or 100%). Once scores close to 100% are attained then no further improvement can be recognised, as the testing materials are not of sufficient difficulty to challenge the patient's abilities. These effects can distort results and make it difficult to reveal significant differences in speech recognition ability (Gifford, Shallop, & Peterson, 2008).

There is less redundant information in single monosyllabic words than there is in sentences, which yield multiple contextual clues involving syntax and semantics. Single word recognition tests are not representative of spoken language and the validity of these word lists for predicting the social adequacy of one's hearing has been questioned (Orchik, Krygier, & Cutts, 1979; Beattie, 1989).

1.3 Speech-in-Noise

As there is no correlation between self-reported measures of difficulty understanding speech-in-noise and objective measurements of this ability (Rowland, Dirks, Dubno, & Bell, 1985), efficient, reliable objective tests should be part of the audiological test battery. Speech-in-noise tests have long been

recognised as an important addition to the audiological test battery, although they are only just starting to be introduced clinically (Carhart & Tillman, 1970; Dirks, Morgan, & Dubno, 1982; Strom, 2006). Speech-in-noise testing enables the clinician to test hearing impaired listeners in the kind of 'real-world' situations in which they report having the greatest difficulty. It can have benefits for hearing aid selection and counselling, giving a more realistic assessment of the likely benefit the patient will receive from hearing aids (Beattie et al., 1997).

Some of the most common speech-in-noise tests are the Connected Sentence Test (CST; Cox, Alexander, & Gilmore, 1987), the Hearing in Noise Test (HINT; Nilsson, Soli, & Sullivan, 1994), the Quick Speech-in-Noise Test (QuickSIN; Killion, Niquette, Gudmundsen, Revit, & Banerjee, 2004), the Bamford-Kowal-Bench Speech-in-Noise Test (BKB-SIN; Niquette et al., 2003; Etymotic Research, 2005), the Words-in-Noise test (WIN; Wilson, 2003; Wilson & Burks, 2005) and the digit triplet test (Smits, Kapteyn, & Houtgast, 2004; Ozimek, Warzybok, & Kutzner, 2010; Zokoll, Wagener, Brand, Buschermöhle, & Kollmeier, 2012). The CST, HINT, QuickSIN and BKB-SIN use sentence level materials as the target stimuli; the WIN uses monosyllabic words, and the digit triplet test uses a sequence of 3 digits.

1.4 Masking Noise

The speech-in-noise tests listed above use multi-talker babble as the masking noise with the exception of the HINT, which uses speech-spectrum noise. Wilson and colleagues (2007) compared the effectiveness of the HINT, QuickSIN, BKB-SIN, and WIN tests in differentiating between speech recognition performance by listeners with normal hearing and performance by listeners with hearing loss. The separation between groups was least with the BKB-SIN and HINT (4–6 dB) and most with the QuickSIN and WIN (8–10 dB). While differences in semantic context contribute to the performance on each test, background masking noise also has an effect. Speech-spectrum noise waveforms like those used in the HINT exhibit little amplitude modulation (AM), whereas, depending on the number of

talkers, multi-talker babble usually has a larger AM characteristic. The importance of an AM characteristic is that during the low point in the waveform fluctuation the signal-to-noise ratio (SNR) is increased, thereby offering the listener a glimpse of a portion of the target speech signal (Miller & Licklider, 1950; Dirks & Bower, 1970; Howard-Jones & Rosen, 1993). Hearing impaired listeners have greater difficulty than normal hearing listeners taking advantage of the momentary improvement in SNR due to their poorer temporal resolution abilities (Stuart & Phillips, 1996, 1998).

One class of masking noise that may be particularly useful in the assessment of speech-in-noise abilities is interrupted noise (Miller, 1947; Miller & Licklider, 1950; Pollack, 1954, 1955; Carhart, Tillman, & Johnson, 1966; Wilson & Carhart, 1969). Interrupted noise is usually a continuous noise that has been multiplied by a square wave that produces alternating intervals of noise and silence. Wilson and Carhart (1969) found that spondaic word thresholds for listeners with normal hearing were 28 dB lower in an interrupted noise than in a continuous noise, whereas listeners with hearing loss experienced only an 11 dB difference. This is a wider separation of recognition performance than is provided by either the QuickSIN or the WIN. The use of amplitude modulated, interrupted noise as the masker may provide a more sensitive measure of hearing impairment than the multi-talker babble currently used in clinically available speech-in-noise tests.

1.5 Adaptive Testing Procedures

Speech tests in quiet of the kind that are currently used in audiological practice in New Zealand fall into the category of non-adaptive tests. A non-adaptive testing procedure, where the distribution of trials is pre-determined at different fixed intensities, is called the method of constant stimuli. The BKB-SIN, QuickSIN and WIN tests use a modified method of constant stimuli in a descending presentation level paradigm, which is a pseudo-adaptive procedure involving the presentation of a set of target stimuli at a fixed SNR followed by further sets of target stimuli at decreasing levels. The number of target stimuli and decibel step sizes can be

varied, but all are administered in a systematic fashion. The Spearman-Kärber equation (Finney, 1952) is used to calculate the SRT of 50%. The HINT uses a truly adaptive procedure (Levitt, 1971) whereby the stimulus level on any one trial is determined by the response to the preceding stimulus. Threshold is defined as the stimulus intensity at which the listener can identify the stimulus correctly for 50% of trials. By measuring the SRT directly, rather than eliciting percentage correct scores, floor and ceiling effects are avoided. Furthermore, by honing in more quickly and efficiently on the region of interest where the individual's threshold is likely to fall, adaptive tests can be more effective than tests that use the method of constant stimuli (Levitt, 1978) while still preserving accuracy and reliability (Buss, Hall, Grose, & Dev, 2001; Leek, 2001). This has important ramifications for clinicians and time management while making the task less onerous for the patient.

1.6 Sentence Tests

Sentences are far more representative of everyday communication than isolated monosyllabic words or digit triplets since they include natural intensity fluctuations, intonation, contextual cues, and temporal elements that are associated with conversational speech (Nilsson et al., 1994). Conversational speech is highly redundant, as knowledge of the subject in question, and visual cues from lip-reading and body language can assist the listener in deciphering the signal. From a measurement point of view, the psychometric functions of sentences are steeper than those of words and digits (McArdle, Wilson, & Burks, 2005), making sentences particularly suitable for accurate estimation of the SRT.

Two types of sentence tests can be distinguished, namely those based on everyday utterances of unified grammatical structure, and those comprising semantically unpredictable sentences of a fixed grammatical structure. The first type of sentence test was originally proposed by Plomp and Mimpen (1979) and the test materials are called Plomp-type sentences. Plomp-type sentence tests have been developed for Dutch (Plomp & Mimpen, 1979; Versfeld, Daalder, Festen, &

Houtgast, 2000), German (Kollmeier & Wesselkamp, 1997), American English (Nilsson et al., 1994), Swedish (Hallgren, Larsby, & Arlinger, 2006), French (Luts, Boon, Wable, & Wouters, 2008) and Polish (Ozimek, Kutzner, Sek, & Wicher, 2009) languages. In general, these tests are composed of phonemically and statistically equivalent lists made up of different sentences, where differences in the phonemic distribution and list-specific SRTs across lists are statistically insignificant. The disadvantage of Plomp-type sentences is that the test lists usually cannot be used twice with the same subject within a certain time interval (i.e., shorter than half a year). The meaningful sentences can easily be memorized or particular words can be guessed from the context, which would affect the SRT result. As the amount of test lists are limited, Plomp-type sentence tests are not suitable when many speech intelligibility measurements have to be performed with the same listener, e.g., during hearing instrument fitting or in research.

The second type of sentence test is the so-called matrix sentence test first proposed by Hagerman (1982) for the Swedish language. These are syntactically fixed, but semantically unpredictable sentences, each consisting of 5 words (name, verb, quantity, adjective, object). There is a base list consisting of 10 sentences with 5 words each. The test sentences are generated by choosing one of the 10 alternatives for each word group in a pseudo-random way that uses each word of the base list exactly once (e.g. Lucy sold twelve cheap shoes). Wagener improved on the idea of Hagerman by taking co-articulation into consideration in order to provide a natural prosody of the synthesized sentences for the German (Wagener, Kühnel, & Kollmeier, 1999a; Wagener, Brand, & Kollmeier, 1999b, 1999c) and Danish (Wagener, Josvassen, & Ardenkjaer, 2003) versions of the matrix sentence test. More recently, British English (Hall, 2006; Hewitt, 2007), Polish (Ozimek et al., 2010) and Spanish (Hochmuth et al., 2012) versions have been developed. The matrix sentence test has become the standard speech audiometry measure in much of Europe, and was selected by the HearCom (www.hearcom.eu) consortium as a means of standardising speech audiometry across the different European regions and languages

1.7 Advantages of Matrix Sentence Tests

Matrix sentence tests have advantages over currently available speech-in-noise tests. A 5 x 10 matrix yields 10^5 or 100,000 different sentence combinations, resulting in a practically unlimited amount of speech material in comparison to Plomp-type sentences. Matrix sentences are useful for hearing aid evaluation and other applications where repeated testing is required. They are also suitable for severely hearing impaired and cochlear implant users because they are spoken relatively slowly and consist of only 50 well known words. The limited vocabulary also makes matrix sentences suitable for testing children.

Unlike Plomp-type sentences, the fixed format of the matrix sentences has the advantage of being very similar across different languages. Measurement and scoring procedures are more uniform making across country comparisons of performance much easier. However, differences do exist between speakers and languages. The British English matrix sentence test (Table 1) would not be appropriate for New Zealand speakers in its original form.

Name	Verb	Numeral	Adjective	Object
Peter	got	three	large	desks
Kathy	sees	nine	small	chairs
Lucy	bought	five	old	shoes
Alan	gives	eight	dark	toys
Rachel	sold	four	thin	spoons
Barry	likes	six	green	mugs
Steven	has	two	cheap	ships
Thomas	kept	ten	pink	rings
Hannah	wins	twelve	red	tins
Nina	wants	some	big	beds

Table 1 - British English word matrix

1.8 New Zealand English

New Zealand English differs from British English in a number of ways, most noticeably in the vowel system. New Zealand English vowels have a very different formant structure and place in the vowel space (Maclagan & Hay, 2007). The front vowels in FLEECE, DRESS and TRAP within the lexical sets introduced by Wells (1982) have raised and fronted in New Zealand English causing them to be pronounced much higher in the mouth, similar to Australian and South African English. DRESS is sometimes pronounced so high in the vowel space by some speakers that it can overlap with FLEECE. There is further neutralisation of front vowels before /l/, such as in the word pairs *celery*/*salary,* and *doll, dole* and *dull*. The vowel in KIT has centralised and lowered even further than when Wells (1982) described it. NEAR and SQUARE are completely merged for many speakers, so that *cheer* and *chair*, *beer* and *bare* are pronounced identically. The GOOSE vowel is also very central, even fronted in some cases, except before /l/. Given that speech perception materials will be presented to hearing impaired individuals under challenging listening conditions, these differences in phonology could have an impact on the performance of New Zealand English speakers on the British English matrix sentence test. For example, "desks" could be confused with "disks" by a New Zealand listener, and hence some substitutions of the words in the British English matrix (Table 1) would be required.

1.9 Auditory-visual Enhancement

A further criticism of speech audiometry in New Zealand is that recorded test material is presented in the auditory-alone condition, which fails to account for the influence of visual cues on speech intelligibility. Evidence suggests that speech intelligibility improves when listeners can both see and hear a talker, compared with listening alone (Sumby & Pollack, 1954; Grant, Walden, & Seitz, 1998). Watching the face of a talker while listening in the presence of background noise can yield an effective improvement of up to 15 dB in the signal to noise ratio relative to auditory-alone (Sumby & Pollack, 1954). Often the advantage of

supplementing listening with watching is more than additive (Sommers, Tye-Murray, & Spehar, 2005). One reason for this superadditive effect is the complementary nature of the auditory and visual speech signals (Grant et al., 1998). For example, cues about nasality and voicing are typically conveyed very well by the auditory signal, even in adverse listening situations, whereas the visual signal does not convey them at all. On the other hand, cues about place of articulation are conveyed by the visual signal but not very well by a degraded auditory signal, as when listening with a hearing loss or listening in the presence of background noise (Tye-Murray, Sommers, & Spehar, 2007).

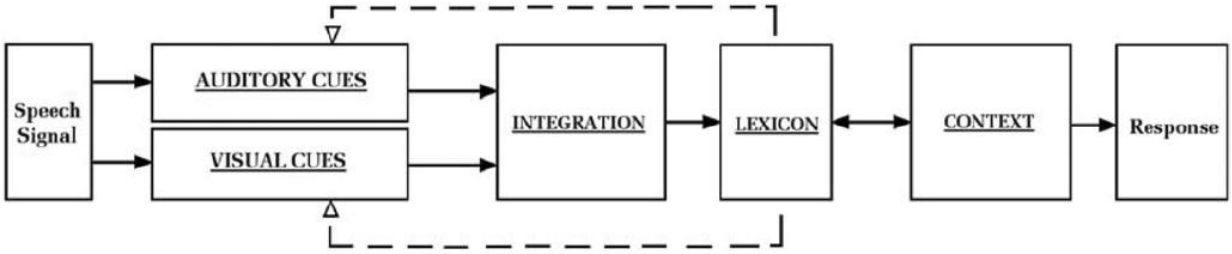


Figure 1 - Schematised framework of auditory-visual speech recognition

Grant et al. (1998) proposed a conceptual framework (Figure 1) for understanding the improved performance for auditory-visual presentations in which both peripheral and central mechanisms contribute to an individual's ability to benefit from combining auditory and visual speech information. In the initial step of the model, peripheral sensory systems (audition and vision) are responsible for extracting signal related segmental and suprasegmental phonetic cues independently from the auditory and visual speech signals. These cues are then integrated and serve as input to more central mechanisms that incorporate semantic and syntactic information to arrive at phonetic and lexical decisions.

Grant et al. (1998) reported great variability among adults with hearing impairment in their ability to integrate auditory and visual information during the process of speech perception. Tye-Murray et al. (2008) found that older adults are less able to use visual cues than younger people. Grant et al. (1998) suggested that poor audio-visual perception of speech may result from difficulties in different

areas such as auditory perception, visual perception, integration ability of the sensory information, ability to use contextual and language information, and working memory. Speech audiometry that presents monosyllabic words in quiet in the auditory-alone modality fails to account for many of these factors. The presentation of sentences in noise in the auditory-visual modality may provide a better measure of real world speech perception.

1.10 Statement of the Problem

Speech audiometry in New Zealand generally utilises monosyllabic words presented in quiet. Speech recognition testing in quiet does not address the main problem experienced by the majority of hearing impaired listeners, which is difficulty understanding speech-in-noise. Furthermore, the method of constants currently used to measure SRTs is susceptible to floor and ceiling effects. Matrix sentence tests have a number of advantages over currently available New Zealand speech tests. The sentence material provides a better representation of everyday spoken language than single words. The potential for 100,000 different sentence combinations gives a practically unlimited set of sentence material, which is useful for repeated testing applications such as hearing aid evaluations. The relatively slow speaking rate and simple vocabulary makes matrix sentences suitable for testing the severely hearing impaired and children. The matrix sentence test utilises masking noise to provide a better assessment of listening abilities in background noise. The use of amplitude modulated, interrupted noise may provide better separation between normal hearing and hearing impaired listeners than the multi-talker babble and speech spectrum noise currently used in clinically available speech-in-noise tests. Matrix sentence tests are compatible with adaptive SRT seeking procedures, which avoid floor and ceiling effects, and are more efficient than the method of constants. A British English version of the matrix sentence test has already been developed. However, differences in phonology between British and New Zealand English may compromise the validity and reliability of the test when used for New Zealand English speakers. Thus, a New Zealand English matrix sentence test needs to be developed. While the method for producing matrix sentence tests in an auditory-alone modality is well documented, the author is unaware of any versions available in an auditory-visual format. A new procedure will therefore be developed to allow the New Zealand English matrix sentence test to operate in auditory-alone, visual-alone or auditory-visual modes. The addition of visual cues to speech audiometry testing may provide a more accurate evaluation of the difficulties experienced by hearing impaired listeners in the real world.

2 Methodology

2.1 Composition of Base Matrix

The British English word matrix (Table 1) was used as the basis for development of the New Zealand English word matrix (Table 2).

Name	Verb	Quantity	Adjective	Object
Amy	bought	two	big	bikes
David	gives	three	cheap	books
Hannah	got	four	dark	coats
Kathy	has	six	good	hats
Oscar	kept	eight	green	mugs
Peter	likes	nine	large	ships
Rachel	sees	ten	new	shirts
Sophie	sold	twelve	old	shoes
Thomas	wants	some	red	spoons
William	wins	those	small	toys

Table 2 - New Zealand English word matrix

The base list consists of 10 different five-word sentences with the same syntactical structure (name, verb, quantity, adjective, object), which is consistent with the format used by other language versions of the matrix sentence test. The composition of the word matrix aims to achieve a balanced number of syllables within word groups, semantic neutrality and grammatical correctness, and to match the language-specific phoneme distribution (Hochmuth et al., 2012).

The boxes in Table 2 highlight the words that differ from the British matrix. Substitution of some of the words in the British matrix was necessary in order to remove potential vowel confusions during open-set testing for speakers of New Zealand English. Substitution of other words were necessary in order to best match the phonemic distribution of New Zealand English.

The substitutions between the British and New Zealand matrices were:
"Amy" replaces "Alan" for gender and phonemic balance
"David" replaces "Barry" for phonemic balance
"Oscar" replaces "Lucy" for gender and phonemic balance
"Sophie" replaces "Steven" for gender and phonemic balance
"William" replaces "Nina" for gender and phonemic balance
"those" replaces "five", which has the same vowel as "nine"
"good" replaces "pink", which may be confused with "punk"
"new" replaces "thin" for phonemic balance
"bikes" replaces "beds", which may be confused with "bids"
"books" replaces "chairs", which may be confused with "cheers"
"coats" replaces "desks", which may be confused with "disks"
"hats" replaces "rings", which may be confused with "rungs"
"skirts" replaces "tins", which may be confused with "tens"

While there is no "gold standard" for the distribution of phonemes in New Zealand English, the phonemic distribution of the New Zealand Hearing in Noise Test (NZHINT; Hope 2010) provided a basis for comparison. The NZHINT is based on five hundred sentences of 5-7 syllables collected from New Zealand children's books and recorded by a native New Zealand English speaker.

The phonemic distribution of the UC Auditory-visual matrix (New Zealand English) was compared against the NZHINT (Figure 2).

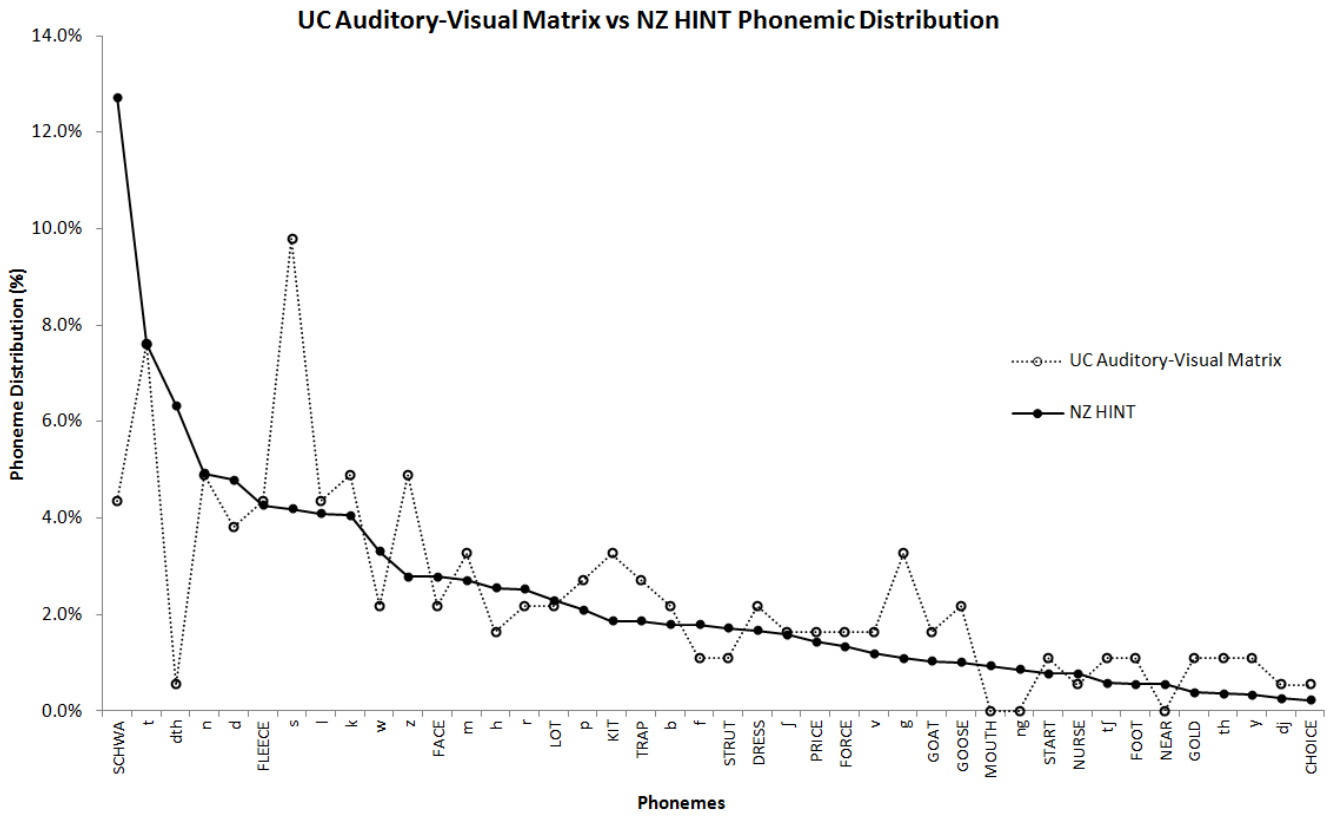


Figure 2 - Phonemic distribution of UC Auditory-visual matrix vs NZHINT

There was a significant positive relationship between the phonemic distributions [$r = 0.6$, $n = 42$, $p < 0.001$]. In comparison to the NZHINT, the UC Auditory-visual matrix has an underrepresentation of "dth" phonemes as it does not contain any articles such as "the". The UC Auditory-visual matrix has an overrepresentation of "s" phonemes as all nouns are plural.

2.2 Sentence Generation

A list of 100 sentences (Appendix A) were selected from the matrix such that each of the 400 unique word pair combinations (e.g. David-bought, bought-three, three-big, big-books) were included in the sentence recording list. The sentences in the recording list were created by following a pattern (Figure 3) of pairing the name, quantity and object in each row with the verbs and adjectives in the adjacent columns.

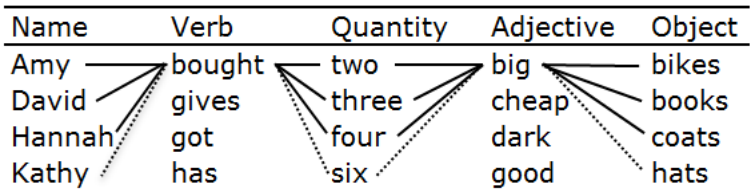


Figure 3 - Matrix sentence generation pattern

In this way, all sentences beginning with the name "Amy" also contained the quantity "two" and the object "bikes". All sentences beginning with the name "David" also contained the quantity "three" and the object "books". Thus, the first three sentences in the recording list were: "Amy bought two big bikes", "David bought three big books" and "Hannah bought four big coats".

2.3 Sentence Recording

A single walled IAC soundproof booth (Industrial Acoustics Company Ltd) was used as a recording studio. The layout of the recording equipment inside the booth is shown in Figure 4.

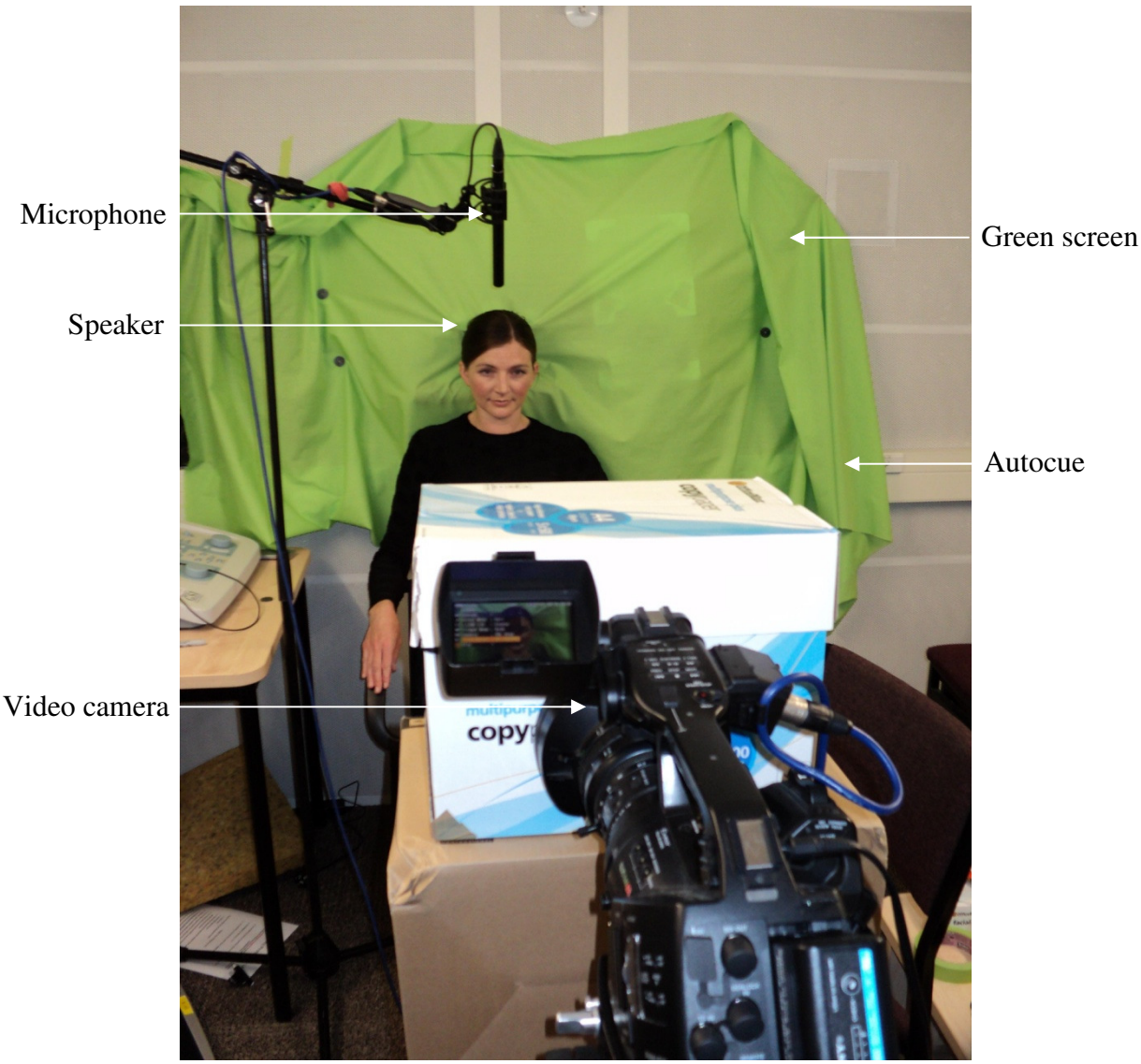
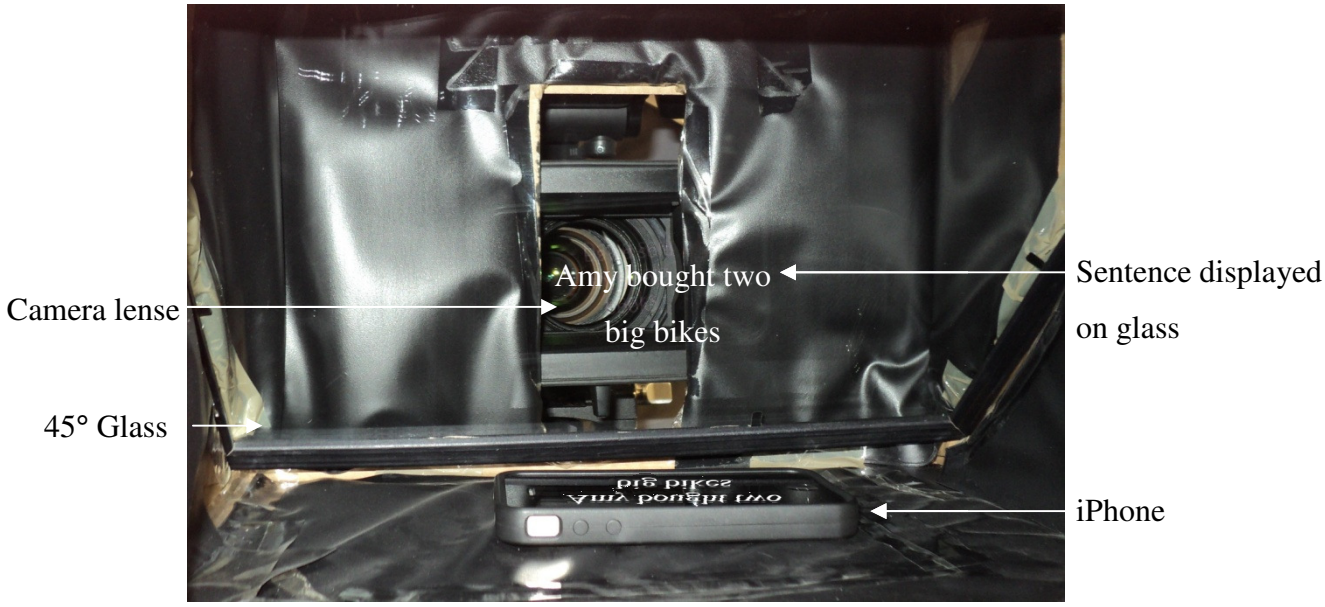


Figure 4 - UC Auditory-visual Matrix Sentence Test recording set-up

An autocue was constructed from a cardboard box, piece of glass and an iPhone (Figure 5).



set-up

The autocue was constructed in a simpler manner than many commercially available autocue systems. A software program was written in LabVIEW (version 9.0.1, National Instruments, TX, USA) that presented the sentence text in a timed manner. Video of this text was captured using Camtasia 6 Studio (TechSmith Corp., MI, USA), which was then inverted and rotated using VirtualDub (www.virtualdub.org). The video was downloaded onto an iPhone which sat in the base of the autocue and projected the text upwards onto a piece of glass angled at 45 degrees. The ghosted image of the text was then able to be read off the front of the glass while the text was invisible to the camera behind the glass due to the angle of reflection. The autocue setup was enshrouded in a cardboard box lined with black plastic to keep the light out and limit reflections from outside the box.

The speaker was a New Zealand born, 32 year old female actress from the School of Fine Arts at the University of Canterbury. The speaker was seated with her back against the wall and her head cradled by a support in order to maintain a stable head position throughout the recording. The head support was covered by a green screen allowing it to be later edited out of the recording. The speaker read the 100 sentences (Appendix A) aloud from the autocue. The sentences were delivered every 3.3 seconds with a 0.9 second gap between each sentence to allow the speaker to return to the mouth shut position. This rate enabled all 100 sentences to be recorded in seven minutes. The recording was captured by a Sony PMW-EX3 video camera and AKG C 568 EB condenser microphone. The video was captured in HD format at a frame rate of 50 fps, pixel resolution of 1280 x 720, and pixel aspect ratio of 1.0 using a progressive scan. The audio was captured in PCM format at a rate of 48,000 Hz. Three recordings of the list of 100 sentences were made in consecutive order with only a 20–30 second gap between each recording. The recordings were saved in mp4 format, and were later transferred to a laptop via USB cable.

2.4 Vowel Recording and Accent Analysis

A sample of the speaker's vowels was taken by recording the speaker voicing the 11 H_D words listed in Table 3. Three sets of these H_D words were recorded on an iPhone and stored in m4a (mpeg4 audio) format.

Lexical Set (Wells, 1982)	H_D Frame	Phoneme (Mitchell, 1946)
Trap	Had	æ
Start	Hard	a
Dress	Head	e
Nurse	Heard	ɜ
Fleece	Heed	¡
Kit	Hid	ɪ
Thought	Hoard	ɔ
Lot	Hod	ɒ
Foot	Hood	ʊ
Strut	Hud	ʌ
Goose	Who'd	u

Table 3 - Vowel notation

The first formant (F1) and second formant (F2) frequencies of the speaker's vowels were analysed using Praat (5.1.32) software. The formant frequencies were averaged over the three recordings (Figure 6).

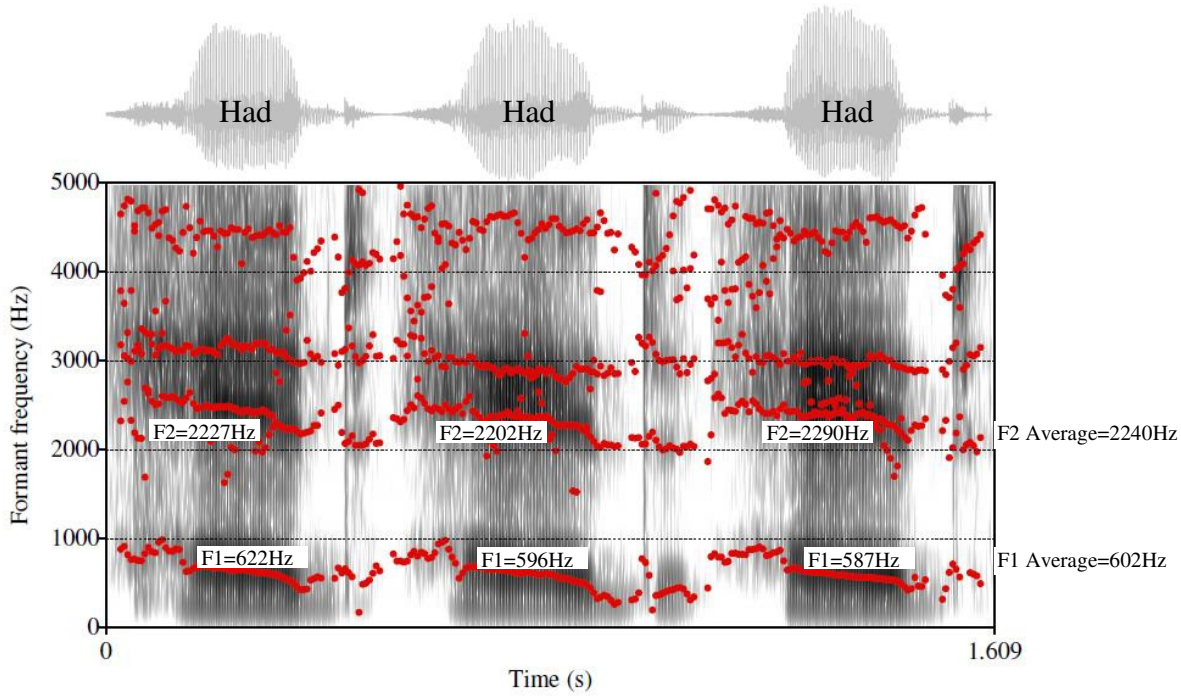


Figure 6 - Formant frequency analysis of the word "Had"

The formant frequencies of the speaker's vowels were compared against normative data from Maclagan and Hay (2007) for typical speakers of New Zealand English of her approximate age (Figure 7). A strong positive correlation was found for F1 [$r = 0.965$, $n = 11$, $p < 0.001$] and F2 [$r = 0.969$, $n = 11$, $p < 0.001$].

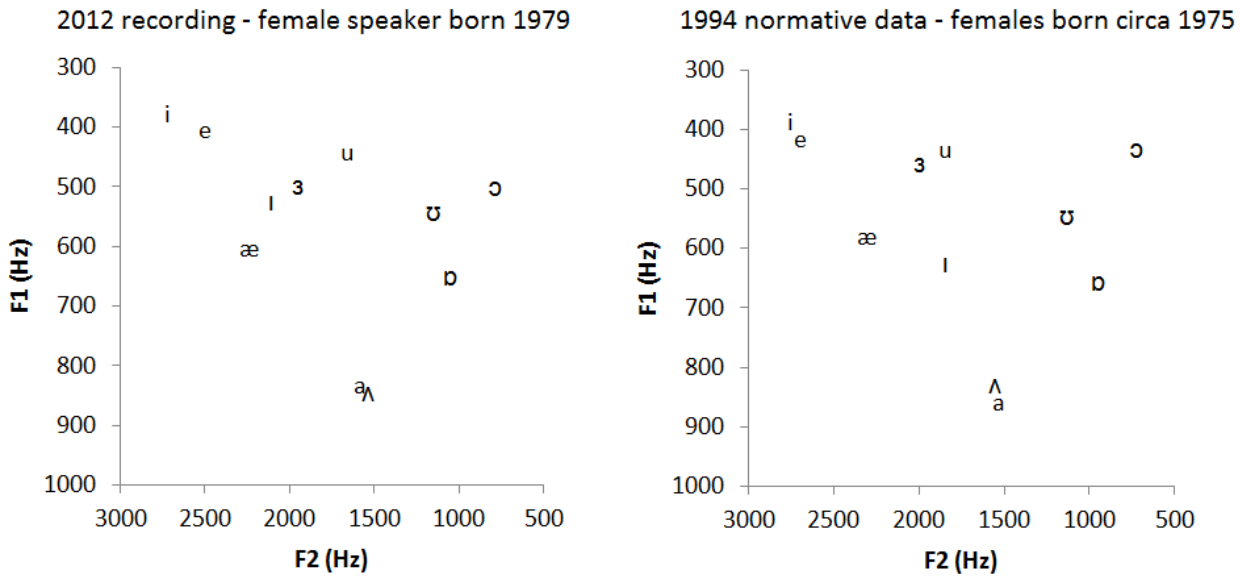


Figure 7 - Speaker's vowel formant frequencies vs normative NZ data

The vowels of New Zealand English are continuously evolving, and there's a wide variety of pronunciations found geographically across the country, but also between generations. The New Zealand accent has become more and more Kiwi[1] with every subsequent generation. UC Associate Professor Margaret Maclagan, an expert New Zealand linguist, judged the speaker's voice to be a typical New Zealand English accent for someone her age. The subjective judgement by an expert listener in combination with the objective analysis of formant frequencies confirmed the speaker had a typical New Zealand accent.

[1] Kiwi is the colloquial name used to describe someone from New Zealand. The name derives from the Kiwi, a flightless bird, which is native to, and the national symbol of, New Zealand.

2.5 Sentence Segmentation

The overall sentence segmentation process is shown below in Figure 8.

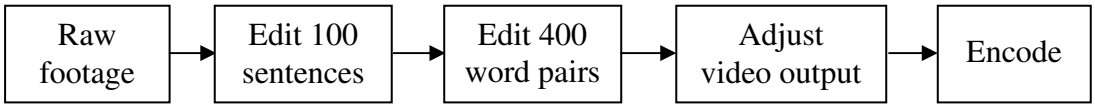


Figure 8 - Auditory-visual sentence segmentation process

Raw Footage

The raw mp4 video file was transferred from the video camera to a Compaq Presario C700 Notebook PC with dual 1.46GHz Intel Pentium processors, 2 GB RAM, running a 32 bit Windows Vista operating system and Adobe Premiere Pro CS4 (V4.2.1) video editing software (abbreviated here as APP). The raw video file was imported into APP in 720p50 format (Table 4).

Audio format	Pulse code modulated (PCM)
Audio sample rate	48,000 samples/second
Video frame size	1280 horizontal x 720 vertical
Video frame rate	50 frames/second
Pixel aspect ratio	1.0
Colour depth	32 bit
Fields	No fields (progressive scan)

Table 4 - 720p50 audio and video format

Edit 100 Sentences

The raw video footage contained all three sentence recording sessions. The final of the three recordings of sentences was selected for segmentation as this was the most consistent in terms of speech and body position (see 3.1 Control of head position). To divide the recording into 100 separate sentences, a cutting point was made at the midpoint of the silence between each sentence (Figure 9).

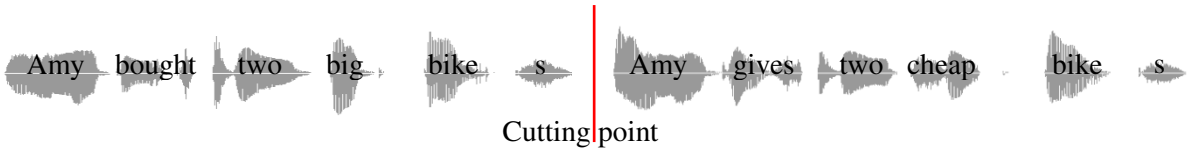


Figure 9 - Auditory-visual segmentation between sentences

The start and end point of the individual sentences was then adjusted. The video was monitored frame by frame to find the point at which the mouth begins to open to form the first word of the sentence (Figure 10). The video was then spooled backwards by 25 frames and a cutting point was made to define the beginning of the sentence.

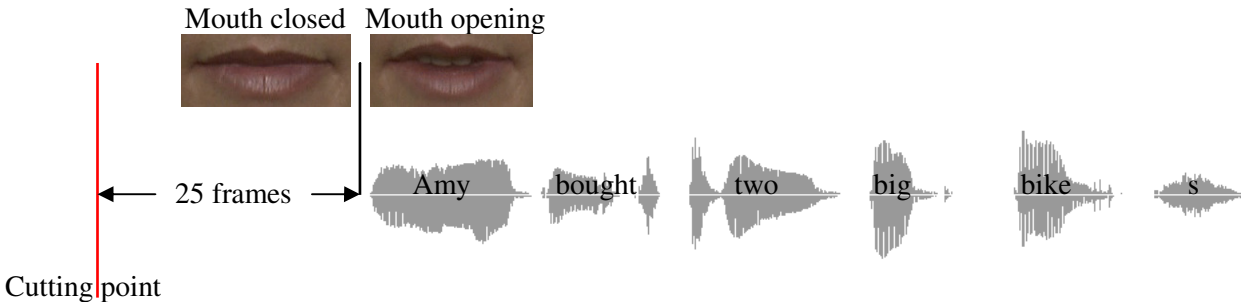


Figure 10 - Auditory-visual segmentation at start of sentence

Similarly, the video was monitored to find the frame where the mouth just closes at the end of the sentence. The video was then advanced another 25 frames and a cutting point was made to define the end of the sentence. This gave consistency to the segmentation procedure such that each sentence contains the same lead in time (0.5 seconds) to the mouth open position, and the same lead out time from the mouth closed position.

Edit 400 Word Pairs

Having cut the first and last words in the sentence according to a fixed lead in / lead out time, the second, third and fourth words in each sentence were cut according to a set of editing rules (Figure 11); the rules were created by trial and error with the objective of finding the smoothest audio and video transition point.

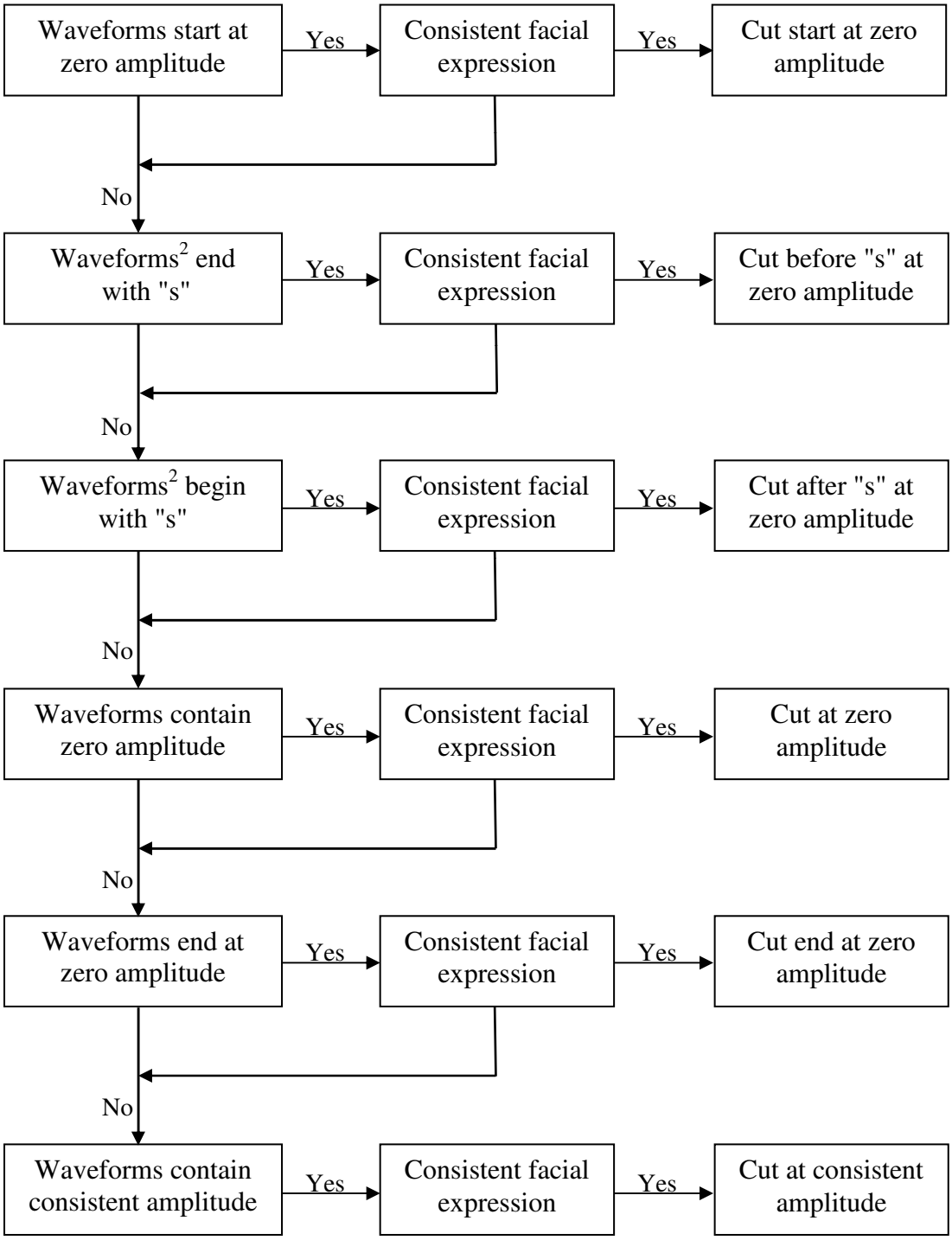


Figure 11 - Auditory-visual sentence segmentation rules

[2] In these cases waveform refers to the words or syllables that begin or end with "s"

The 100 sentences contained 10 instances of each word in the matrix. APP was used to group together each of the 10 instances and inspect the waveforms and video frames for potential segmentation points. The same editing rule was applied to each of the 10 instances.

Waveforms starting at zero amplitude

Appendix 4 shows that all sentences containing the word "bought" have a clearly defined point of zero amplitude at the start of the word. The first two instances are illustrated in Figure 12.

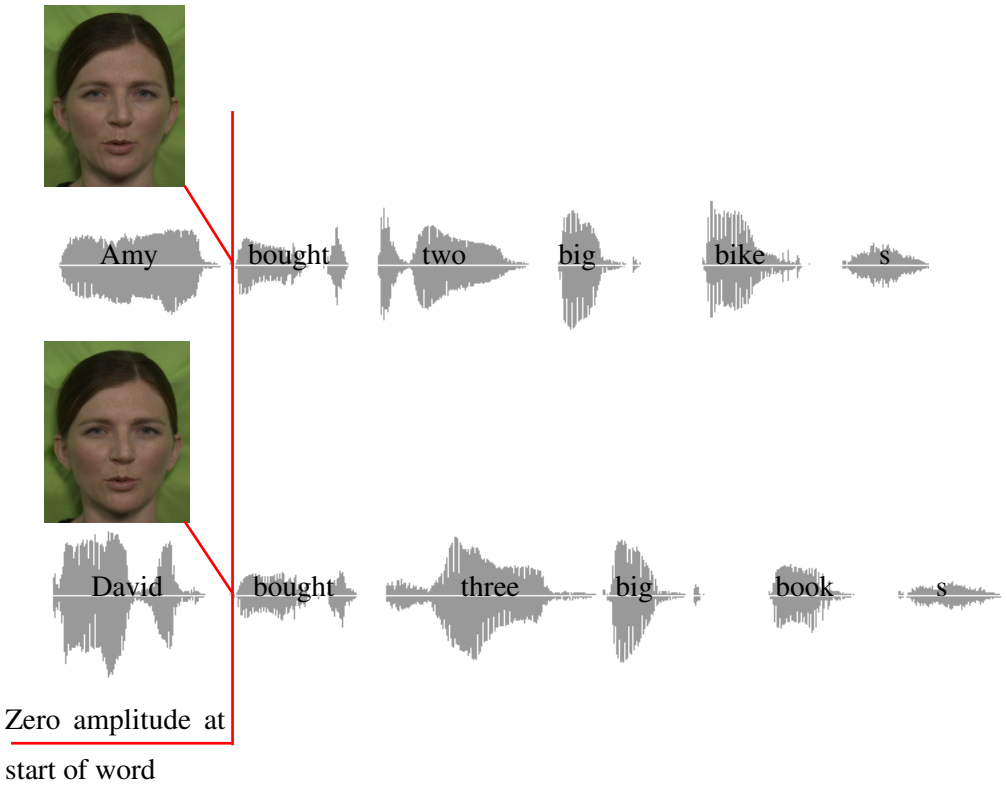


Figure 12 - Auditory-visual segmentation of waveforms starting at zero amplitude

The point of zero amplitude on a waveform represents a silent space between words, which was found to be the most ideal segmentation point for the audio. The video frame at the segmentation point was also inspected to ensure there was a uniform facial expression across all instances of the word. The mouth closed

position was found to be the most ideal segmentation point for the video as it provides a smoother transition between video frames compared to the mouth open position.

Waveforms ending with "s"

It was found that the point of zero amplitude at the start of a word is not always the most ideal segmentation point. Appendix 4 illustrates that words ending with "s" have a point of zero amplitude at the beginning of the "s" waveform. The first two instances of the word "gives" are illustrated in Figure 13.

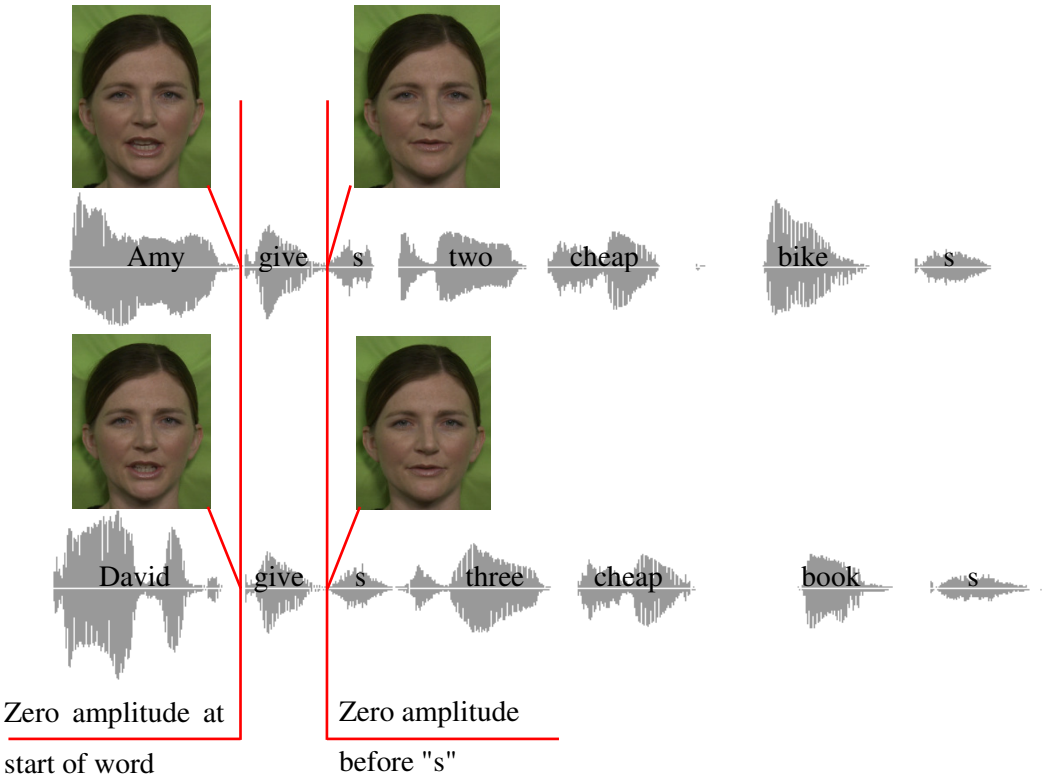


Figure 13 - Auditory-visual segmentation of waveforms ending with "s"

While a point of zero amplitude could be found at the start of each instance of the word "gives", the open mouth position in the video frames was not ideal. Instead, the segmentation point was made before the "s", where the corresponding video frames had a more consistent facial expression and a closed mouth position.

Waveforms beginning with "s"

Appendix 4 illustrates that words beginning with "s" often do not have a point of zero amplitude at the start of the word. Two instances containing the word "some" are illustrated in Figure 14.

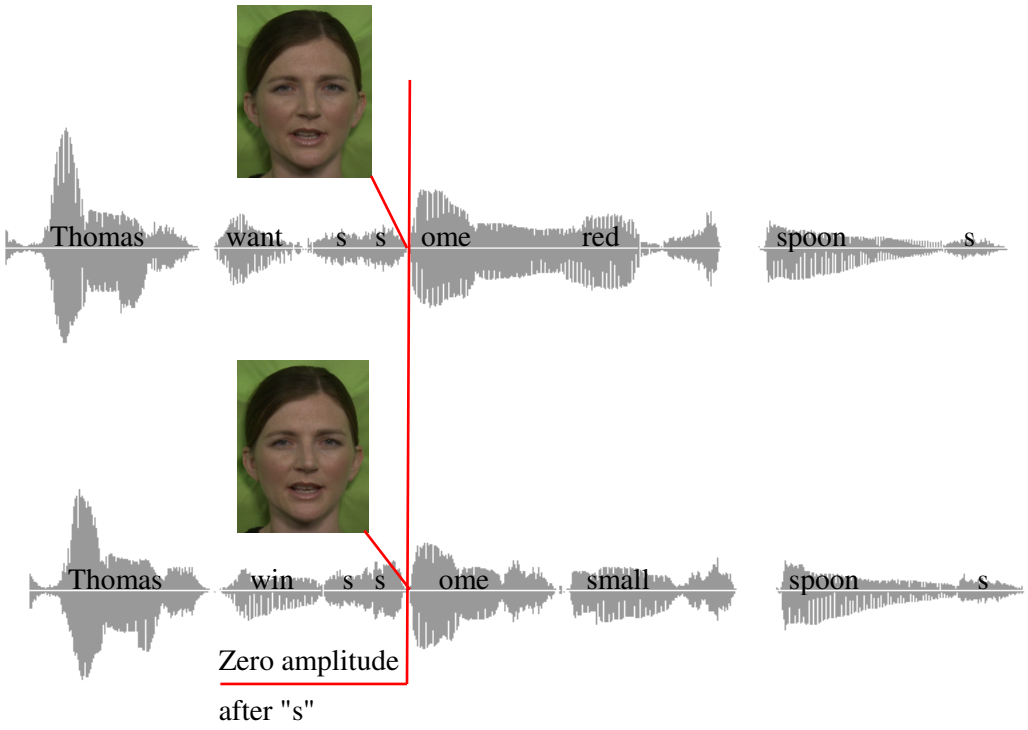


Figure 14 - Auditory-visual segmentation of waveforms beginning with "s"

The "s" waveforms at the end of "wants" and "wins" blend together with the "s" at the start of "some" to eliminate a point of zero amplitude between the words. The most ideal[3] cutting point for the audio was found to be the point of zero amplitude at the end of the "s" waveform. While the corresponding mouth position in many cases was not closed, the selection of an ideal cutting point for the audio took precedence over the video.

[3] most ideal refers to the subjective smoothness of audio and video transition points.

Waveforms containing a point of zero amplitude

Words that did not start or end with "s" were inspected for a point of zero amplitude as a potential cutting point. Two instances containing the word "large" are illustrated in Figure 15.

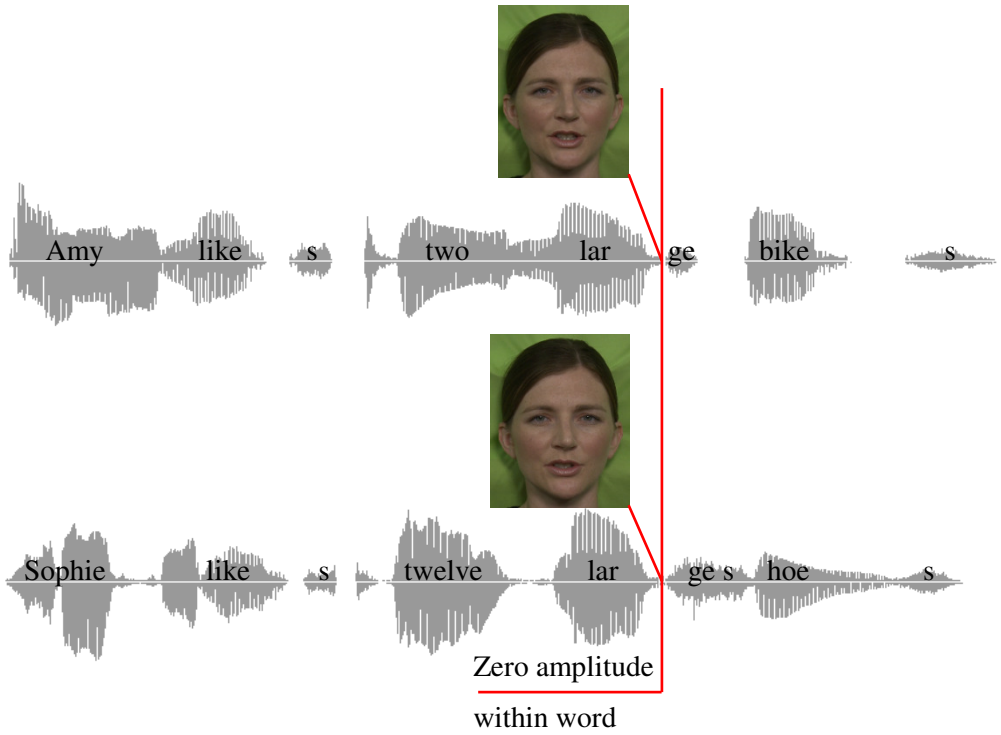


Figure 15 - Auditory-visual segmentation of waveforms containing zero amplitude

It was not possible to make a cut at the start or the end of the word "large" as some instances (e.g. "two large", "large shoes") were blended with the preceding or following waveform, which eliminated the point of zero amplitude between the words. All instances of the word "large" were found to have a point of zero amplitude in the middle of the waveform, which was an ideal cutting point for the audio. With the mouth partially open, it was not the most ideal cutting point for the video, however, the audio cutting point took precedence.

Waveforms ending at zero amplitude

Words that did not contain a point of zero amplitude at the start or within the word were inspected for a point of zero amplitude at the end of the word as a potential cutting point. Two instances containing the word "red" are illustrated in Figure 16.

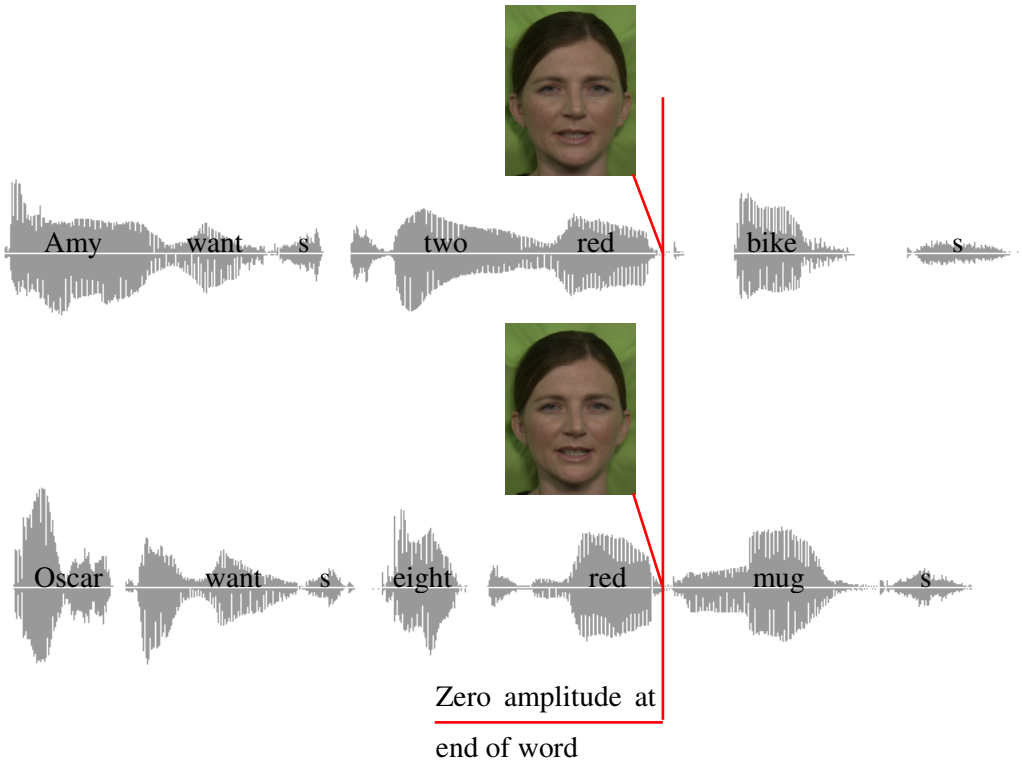


Figure 16 - Auditory-visual segmentation of waveforms ending at zero amplitude

It was not possible to make a cut at the start of the word "red" as some instances (e.g. "two red") were blended with the preceding waveform, which eliminated the point of zero amplitude between the words. The word "red" does not contain a point of zero amplitude within the word. All instances of the word "red" were found to have a point of zero amplitude at the end of the waveform. While the end of a waveform is not usually the most ideal place to make a cut, as it may contain the co-articulation to the next word, the point of zero amplitude was still found to be the best cutting point in terms of the subjective smoothness of the audio and video transition points.

Waveforms containing a point of consistent amplitude

Finally, a point of zero amplitude may not exist at the beginning, in the middle, or at the end of a word. In this case, the waveform was inspected for a point of consistent amplitude[4]. Two instances containing the word "new" are illustrated in Figure 17.

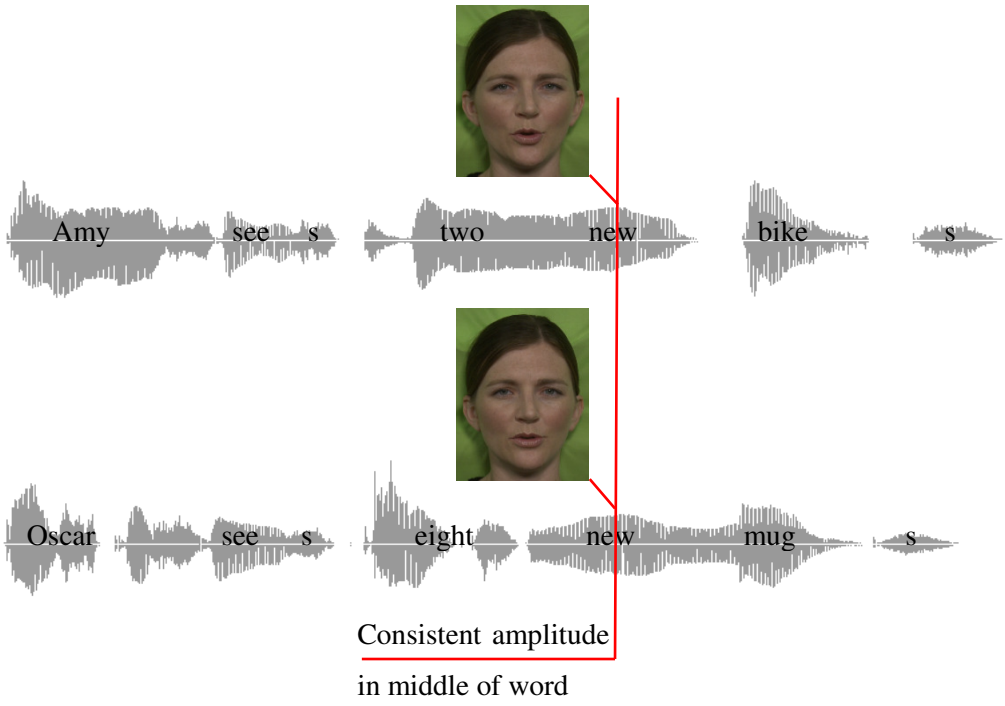


Figure 17 - Auditory-visual segmentation of waveforms containing consistent amplitude

It was not possible to make a cut at the start of the word "new" as some instances (e.g. "two new") were blended with the preceding waveform. It was not possible to make a cut at the end of the word "new" as some instances (e.g. "new mugs") were blended with the following waveform. The word "new" does not contain a point of zero amplitude within the word. While a point of non-zero amplitude is not usually the most ideal place to make a cut, as it may create an audible "transient", cutting at a point of consistent amplitude was found to be an acceptable compromise. As there were numerous options for selecting a point of

[4] Consistent amplitude refers to a point where the amplitude of two separate waveforms is approximately equal.

consistent amplitude, the point where their mouth position was closed was chosen in order to provide the best video cutting point.

Adjust Video Output

The video output was adjusted (Figure 18) in order center the viewer's attention on the speaker's face. The brightness and contrast of video was adjusted to better illuminate the speaker's facial features. The head support system (see 3.1 Control of head position), which was used to maintain the speaker's head in a constant position throughout the recording, created a noticeable pattern of creases on the green screen. In order to remove this, a chroma key was applied to the video output, which replaced the green background with a black background. The chroma key relies on being able to distinguish green from the other colours in the video. The chroma key was unable to remove all of the dark green shadow on the speaker's right side, leaving a patch of green fuzz next to the speaker's neck. In order to remove this, a garbage matte was applied, which blocks anything outside of the matte from appearing in the final video output. The video adjustment procedure was applied to all 400 word pair segments.

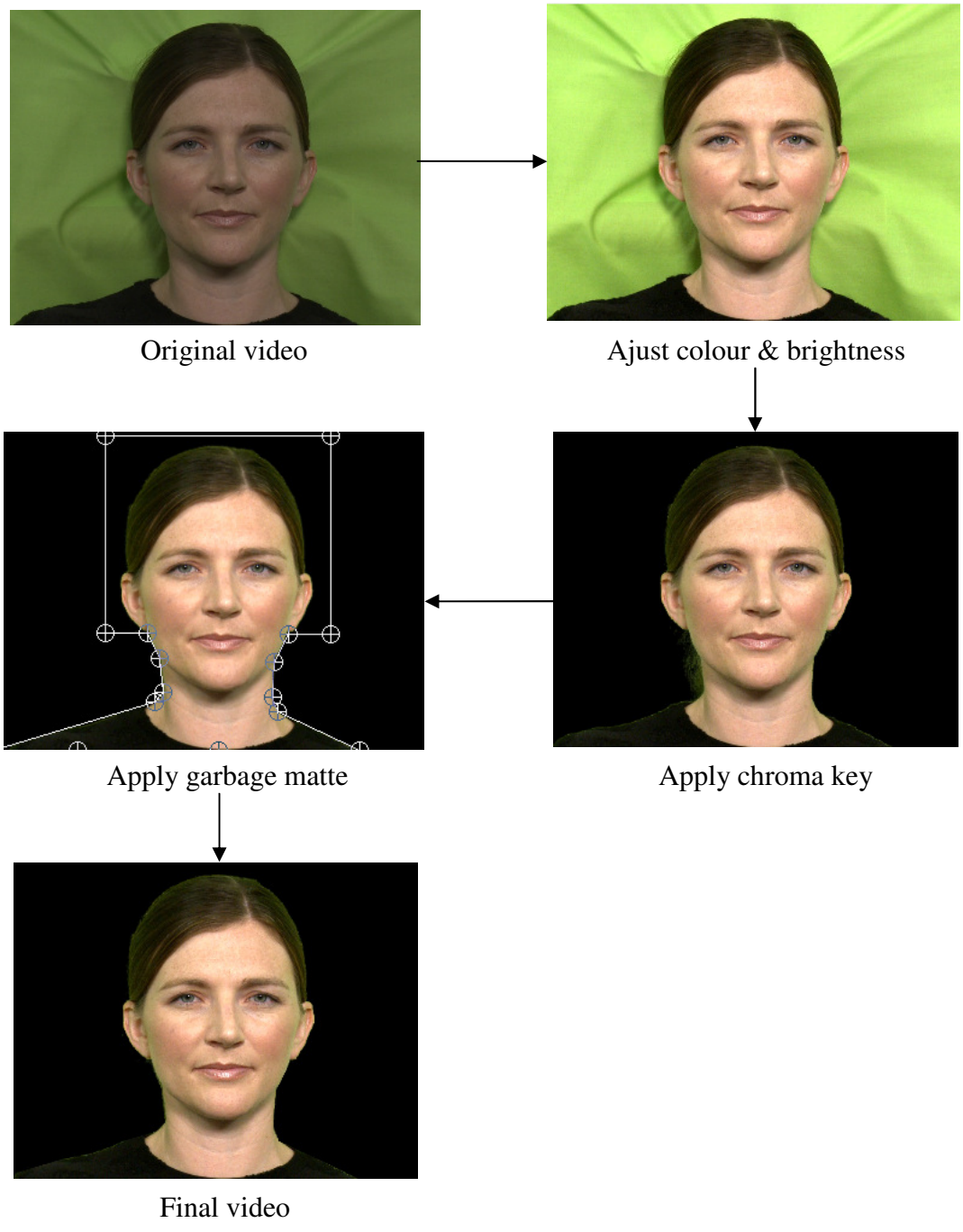


Figure 18 - Post recording adjustment of video output

Encode

Encoding is the process of writing audio and video files to a format suitable for presentation to the end user. The overall encoding process is illustrated in Figure 19.

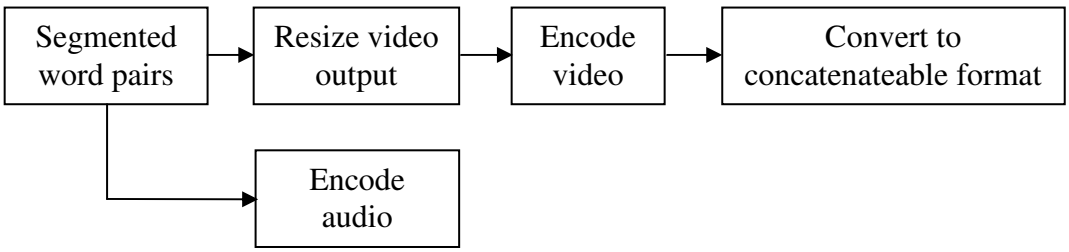


Figure 19 - UC Auditory-visual Matrix Sentence Test encoding process

Segmented word pairs

Adobe Media Encoder CS4 Version 4.2.0.2006 (abbreviated here as AME) was used to extract the 400 segmented word pairs from APP and separate them into their audio and video components.

Resize video output

Throughout the segmentation process the video was maintained in its original high definition 720p50 format (Table 4). Starting from high definition, the video can be compacted to suit many user interface displays such as computer and television screens without having to modify the sentence segmentation points. The user interface for the project was a computer screen with a small video player window measuring 640 horizontal x 480 vertical pixels. AME was used to separate the audio and video portions of the word pair segments into individual audio and video files (e.g. amy_bought.wav, amy_bought.avi).

The video output was resized to fit the user interface using the following parameters:

Video format	Uncompressed Microsoft avi
Video frame size	640 horizontal x 480 vertical
Video frame rate	50 frames/second
Pixel aspect ratio	1.0
Colour depth	32 bit
Fields	No fields (progressive scan)

Table 5 - UC Auditory-visual Matrix Sentence Test video output settings

The video files were maintained in an uncompressed format in order to preserve their high definition. The video files were outputted to a 250 GB external hard drive (TEAC HD3U S/N 7201188) for storage as each of the 400 uncompressed video files was approximately 200 MB in size. The video files were named according to the word pair they represent e.g. amy_bought.avi.

Encode video

FFmpeg (www.ffmpeg.org) is a free, open source software tool used to record, convert and stream audio and video. FFmpeg contains a library of encoding and decoding software (codecs) for converting audio and video files between formats, which is widely used in the multimedia industry (Cheng, Liu, Zhu, Zhao, & Li, 2011). FFmpeg version SVN-r18709 was used to encode the word pair video files into Microsoft mpeg4 (Moving Picture Experts Group, standard 4) file format, such that they could be played by the standard version of Windows Media Player installed on every Windows based computer. As FFmpeg operates from a MS-DOS (Microsoft Disk Operating System) command line, a batch file was created to automate the process of encoding the 400 word pair video files into mpeg4 format (see Appendix 5 for syntax).

Convert to concatenateable format

The word pair segments needed to be able to be joined together in series to form a sentence (concatenateable). As the mpeg4 video format does not allow concatenation, the word pair video files were converted to mpg (Moving Picture Experts Group, standard 1) format, which does support concatenation. FFmpeg was used to convert from mpeg4 format to a mpg of equal video quality. A batch file was used to automate the process of converting the 400 word pair video files into mpg format (see Appendix 5 for syntax).

Encode audio

AME was used to extract the audio portion of the segmented word pairs from APP and encode them with the following parameters:

Audio format	Windows waveform wav
Audio codec	Pulse code modulated (PCM)
Sample rate	44,100 Hz
Sample type	16 bit
Channels	Mono

Table 6 - UC Auditory-visual Matrix Sentence Test audio output settings

The audio files were named according to the word pair they represent e.g. amy_bought.wav.

2.6 User Interface Development

A user interface (Figure 20) was developed using the LabVIEW (version 9.0.1) development environment. A virtual instrument (VI)[5] was written that mixes the word pair segments together on the fly to produce a seamless sentence in less than 1.5 seconds. Sentences can be presented in auditory-alone, visual-alone or auditory-visual mode. Auditory-alone plays sound without video; visual-alone plays video without sound; while auditory-visual plays sound and video together.

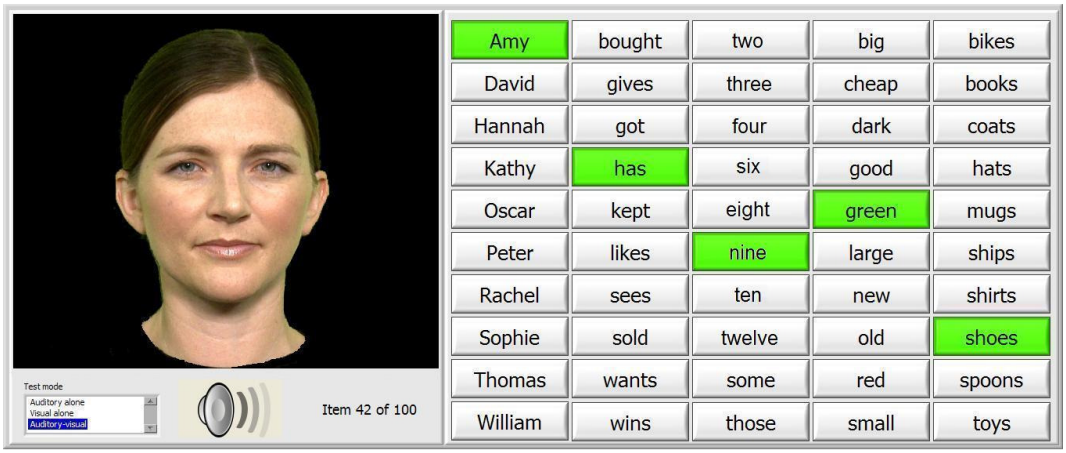


Figure 20 - UC Auditory-visual Matrix Sentence Test user interface

The software code behind the user interface performs three essential functions. The first is to concatenate the audio and video files together to form a sentence. The second is to play the sentence to the user in auditory-alone, visual-alone or auditory-visual mode. The third is to score the response and store the data for analysis. It is essential that each function is performed sequentially.

[5] A virtual instrument, or VI, is the name given to any piece of software written using LabVIEW

Concatenate audio and video files

The software design for performing the concatenation process, which joins the audio and video portions of the word pairs together, is illustrated in Figure 21.

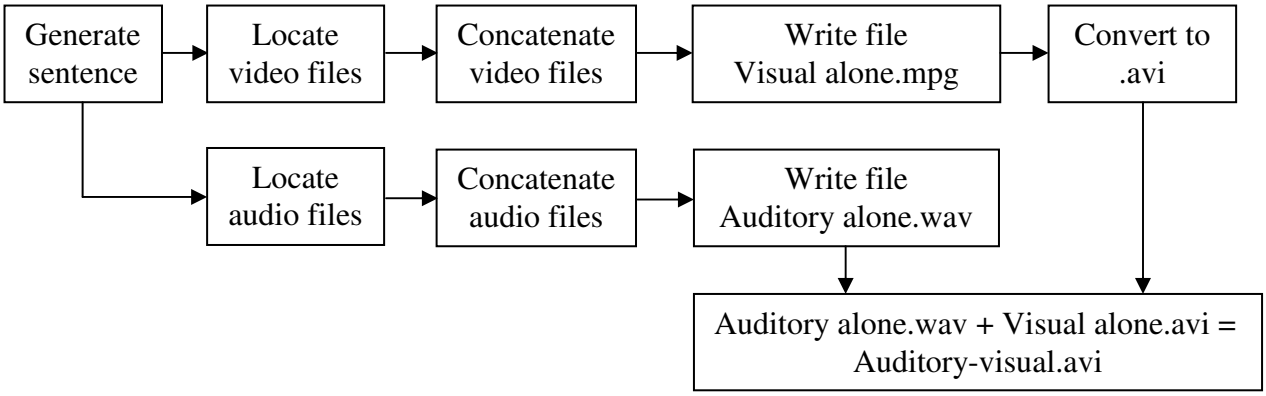


Figure 21 - UC Auditory-visual Matrix Sentence Test mixing software flow chart

Generate sentence

For testing purposes, the developer was able to use the response interface to manually select a sentence to be played by pressing the button of each word. The selected sentence was highlighted in green. In the fully developed version (see 3.8 Future Development), sentences will be generated randomly, or pulled from pre-determined lists.

Locate audio and video files

Based on sentence selected, the software identified the audio and video files that were required to generate the sentence. For example, "Amy bought two big bikes" required the following files:

Audio	Video
amy_bought.wav	amy_bought.mpg
bought_two.wav	bought_two.mpg
two_big.wav	two_big.mpg
big_bikes.wav	big_bikes.mpg

Concatenate audio files

Different methods were required for concatenating the audio files and video files together. LabVIEW was able to treat audio files as arrays of sample values (44,100 Hz, 16 bit, mono) which could then be concatenated to form a longer audio file.

Write audio files

The four word pair audio files making up a sentence were concatenated to form a single wav file. In the case of "Amy bought two big bikes" the word pairs were:

amy_bought.wav + bought_two.wav + two_big.wav + big_bikes.wav

= Auditory alone.wav

The Auditory alone.wav file is a temporarily stored file. It always maintains the same name; however, its content is overwritten each time a new sentence is generated.

Concatenate video files

The standard version of LabVIEW does not include any pre-programmed code for concatenating video files; however, it does allow for an MS-DOS command line to be executed. The mpg video files were joined together by utilising the concatenate function of MS-DOS (see Appendix 6 for syntax)

Write video files

The four word pair video files making up a sentence were concatenated together to form a single mpg file. In the case of "Amy bought two big bikes" the word pairs were:

amy_bought.mpg + bought_two.mpg + two_big.mpg + big_bikes.mpg

= Visual alone.mpg

Visual alone.mpg is also a temporarily stored file, with its content overwritten each time a new sentence is generated. Some attributes critical for playback, such as the length of the file, are not preserved during the concatenation process.

Windows Media Player requires such information in order to playback the file correctly. Visual alone.mpg was re-encoded into mpeg4 format to enable playback by Windows Media Player. The conversion from Visual alone.mpg to Visual alone.avi was made using FFmpeg.

Finally, the Auditory alone.wav and Visual alone.avi files were joined together using FFmpeg to form Auditory-visual.avi (see Appendix 5 for syntax).

Playback sentence

The software design for performing sentence playback is illustrated in Figure 22. Before the playback code can execute, the mixing code must have completed writing the audio and video files Auditory alone.wav, Visual alone.avi and Auditory-visual.avi.

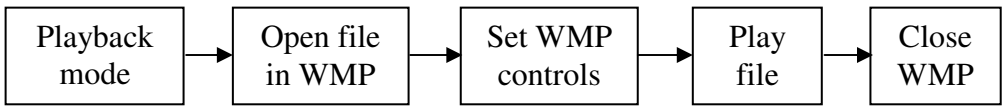


Figure 22 - UC Auditory-visual Matrix Sentence Test playback software flow chart

Playback mode

The user selects the playback mode via the Auditory-alone, Visual-alone and Auditory-visual selection menu in the user interface (Figure 20). Based on that choice, the software selects the corresponding file for playback.

Open file in Windows Media Player

The LabVIEW development environment is able to utilise the functionality of Microsoft Windows Media Player via an Active X container. Active X allows software manufacturers to embed each others code in their applications. The software creates an Active X container for Windows Media Player and then loads either Auditory alone.wav, Visual alone.avi or Auditory-visual.avi.

Set Windows Media Player controls

Windows Media Player controls were set to play the audio and video files in a 640 horizontal x 480 vertical pixel window in the user interface. The standard Windows Media Player controls such as stop, start and volume were hidden from the user so they could not be modified. Control of these variables was instead performed by the software.

Play file then close Windows Media Player

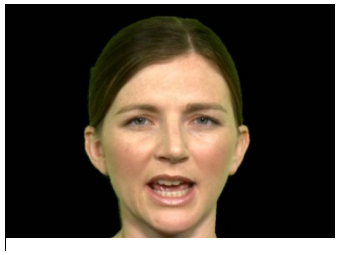
The selected audio and video files play to the end and then close, which allows for the next file to be played.

2.7 Video Transition Analysis

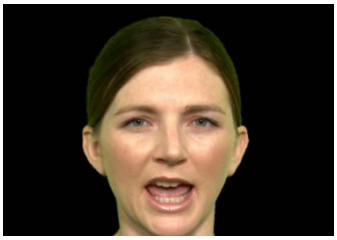
The key feature of matrix sentences is the ability to generate 100,000 unique sentences from just 100 recorded sentences edited into 400 word pairs. However, the large number of possible word pair combinations can result in the pairing of video frames that do not match up sufficiently well enough to produce a smooth looking transition between frames. A method for objectively evaluating the smoothness of the video transitions was developed. A VI was written to extract the first and last frame of each word pair video and store them as jpeg (Joint Photographic Experts Group) images (Figure 23).

nine good (last frame)

good ships (first frame)



sold eight (last frame)



eight big (first frame)

Figure 23 - Audio-visual word pair video transitions

Another VI was written to measure the absolute difference in RGB (Red, Green, Blue) colour channels between images. The maximum possible difference is 78643200 (640 horizontal pixels x 480 vertical pixels x 256 colours). The smaller the absolute difference between images, the smoother the transition. The best transition was "nine-good ships" with a difference of 268393 or 0.34% (268393/78643200 x 100%), while the worst transition was "sold-eight big" with a difference of 1313803 or 1.67%. An analysis was also made of the difference

between images with the mouth region excluded (range = 0.23% - 0.96%). The 400 word pair segments provided 3000 unique transitions. The smoothness of the transitions was normalised by ranking the absolute percentage difference between frames from 1 to 3000 (Figure 24).

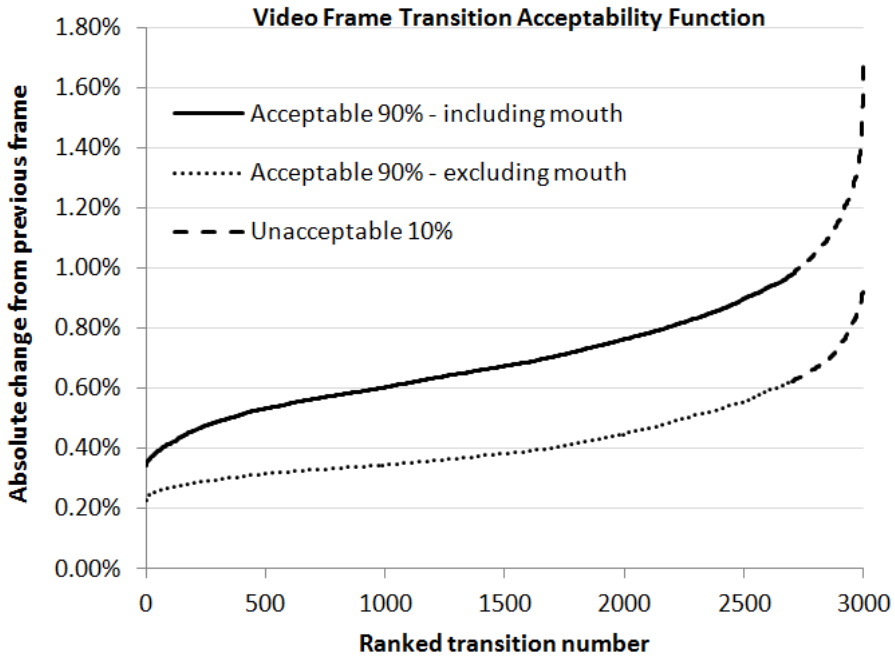


Figure 24 - Normalised smoothness ranking of video transitions

The 3000 video transitions were assigned a quality rating between 0% (worst transition) and 100% (best transition). The lower quality video transitions (e.g. bottom 10%) can be excluded from the matrix sentences; however, the more word pairs that are excluded, the less unique matrix sentences are available. A spreadsheet macro was written to list all 100,000 unique sentences. A lookup function was then used to mark sentences containing any of the unacceptable word pair transitions. The unacceptable sentences were filtered out of the list and the remaining acceptable sentences were counted (Figure 25). Including or excluding the mouth region from the analysis gives essentially the same function.

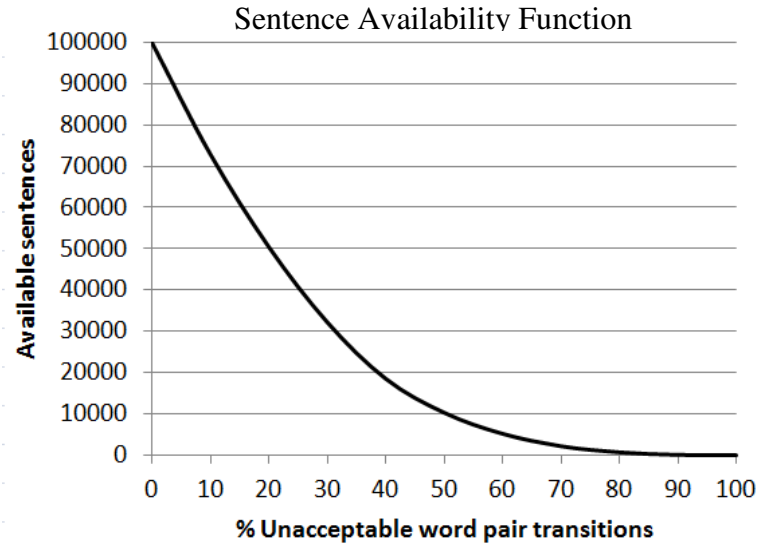


Figure 25 - Percentage of word pairs excluded vs number of available matrix sentences

3 Discussion

3.1 Control of Head Position

A stable head position was found to be essential for the visual component of the UC Auditory-visual Matrix Sentence Test. A number of different schemes were trialled before arriving at the method described in section 2.3.

Without any head support

A trial recording was made of the author voicing the 100 sentences listed in Appendix A without any head support. The sentences were pre-recorded and then played back with a gap between sentences such that the author could listen to the sentence, voice the sentence during the gap, and then return to the mouth closed position in anticipation of the next sentence. Upon editing the recorded sentences into word pairs and then recombining them into new sentences, a number of noticeable jumps were observed in the video output. The jumps were occurring at the cutting points between word pairs. Slight changes in head position throughout the recording were causing a mismatch between video frames when the word pairs were recombined to form new sentences. While the transition of the audio component had a natural sound, the rapid jumps in head position made the video component look unnatural.

Support behind the head

A basic head support system (Head support 1, Figure 26) was constructed behind the green screen of the recording set-up (Figure 4). A piece of Styrofoam was used to provide a basic cradle for the back of the head while the foam and towel provided cushioning. The head support was fixed to the wall with tape in order to hold it in a constant position.

Figure 26 - UC Auditory-visual Matrix Sentence Test head support system

The author made another recording of the 100 sentences (Appendix A). While the head position was more stable than without the head support, the jumps between video editing points was still very noticeable. Changes in facial expression were also noticeable. The first few sentences were made with a lot of enthusiasm and expressiveness; however, as the recording session progressed, tiredness set in and the last few sentences did not have the same energy and facial expression as the first. One of the main factors contributing to the tiredness was trying to maintain the body in a still, upright position for an extended period of time.

Lying on the floor

A trial recording was made with the author lying on the floor with supports on both sides of the head. The lying position reduced the stress on the muscles and required very little effort to maintain position. The resulting word pair video transition points were very stable; however, the affect of gravity pushing straight down on the face produced a distorted, unnatural looking facial expression.

Self correction of head position

A trial recording was made using a visual feedback system. The output of the video camera (Figure 4) was fed into a screen located above the video camera. A sheet of transparent plastic was placed over the screen and the silhouette of the author's head was traced with a marker pen. This allowed the author to monitor his own head position on the screen. Unfortunately, this method of self correction did not prove to be successful. The combination of voicing the sentences, monitoring head position on screen, and making subtle adjustments simultaneously was too much. The feedback screen provided a mirror image, which meant that if the head was out of position to one side, the author needed to move in a counter intuitive direction to correct it. Also, with the feedback screen mounted above the camera lens, the author appeared to be gazing upwards in the recorded video.

Multiple recordings with support behind the head

Experimentation into different head support systems had been undertaken by the author in order to find the best solution before a recording was made by the official speaker. With the support behind the head proving most successful, the method was refined to provide to most stable head position for the speaker. A new head support system (Head support 2, Figure 26) was made from pieces of cardboard in order to provide a more rigid support system with more pressure points in contact with the back of the head. The support was adjusted to the speaker's height while allowing her to sit comfortably in a chair with good back support. The autocue system allowed the speaker to look directly at the camera lens, which avoided previous issues with gaze direction. The autocue system also allowed for a faster presentation of sentences. Reading sentences from an autocue was faster than the auditory presentation system used previously whereby the speaker had to wait for each sentence to be played before repeating the sentence. A faster presentation of sentences reduced the recording time and hence the speaker was not as tired by the end of the recording session. There was still a noticeable difference between the first few sentences and the last few sentences in

the recording. The speaker's facial expression tended to become more sombre and her position would slump further down as the recording progressed. For this reason, three recordings were made in succession without giving the speaker a break in between. In this way, the first recording showed a difference in position and facial expression; while during the second recording the speaker fell into a consistent rhythm, and by the third recording the speech, position and facial expression were quite uniform from start to finish. While the method developed certainly proves the concept of an auditory-visual matrix sentence test can work, there is still room for improvement, especially in the development of better head support systems (see 3.8 Future Development).

3.2 Video Recording Procedures

A number of different video recording techniques were trialled before arriving at the method described in section 2.3.

Video frame rate

Video recordings were initially made using the PAL (Phase Alternating Line) television format, which has been used in New Zealand and around the world for decades as a standard format for analogue television transmission (Arnold, Frater, & Pickering, 2007). The PAL format is captured at a rate of 25 video frames per second. When applied to the UC Auditory-visual Matrix Sentence Test, a frame rate of 25 fps was found to be too slow to capture the changes in mouth position between word pairs. It was found that the mouth position could pass from open to closed within 1 frame, which sometimes caused a misalignment of video frames between adjoining word pairs. This resulted in a noticeable jump in mouth position when sentences were played back.

The 720p50 format (Table 4) was used in the final recording of the UC Auditory-visual Matrix Sentence Test as it has a frame rate of 50 fps, which is double that of the PAL format. The higher frame rate provided better definition to the mouth position, thus making for a smoother transition between word pairs when the sentences were played back. A frame rate greater than 50 fps may provide even better performance; however, there is the issue of video playback compatibility to consider. The 720p50 format is a high definition digital television format that is supported by many standard computer media players such as Microsoft Windows Media Player. Frame rates greater than 50 fps may not be supported by standard media players.

Video recording setup

Section 2.5 describes a number of adjustments that were made to the video output post recording. These adjustments included changes to brightness and contrast and the keying out of the green background. Some of these adjustments were quite time consuming and could have been avoided if more attention had been paid to the recording setup. Figure 27 illustrates a frame from the original recording from which potential improvements can be identified.

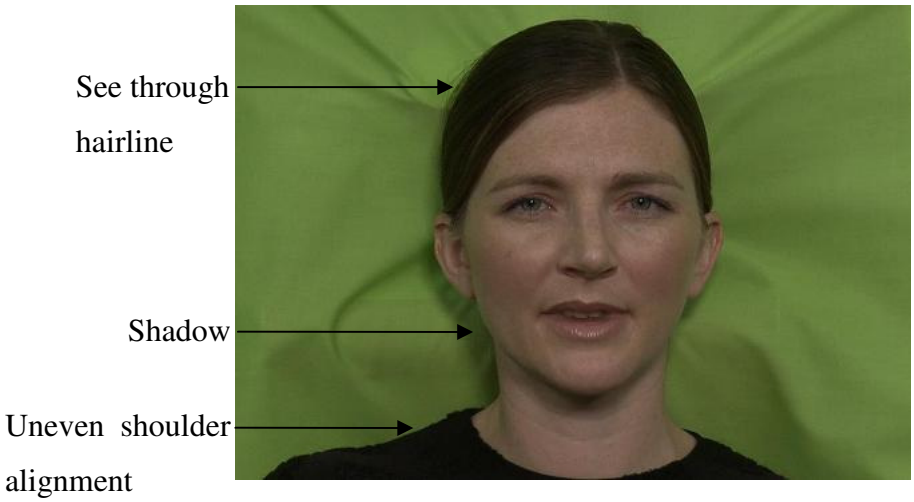


Figure 27 - UC Auditory-visual Matrix Sentence Test recording setup errors

The lighting used was perhaps not bright enough. This resulted in the need to adjust the brightness and contrast of the final video. The lighting was angled such that it created a shadow on the speaker's right side. This shadow, coupled with the crinkles in the green screen from the underlying head support, created a background colour with varying shades of green. The keying process used to remove the green background relies on the difference in colour between the background and foreground. The dark green shadow contained elements of colour similar to the darkness of the speaker's hair, eyes and shirt, which made the separation of background from foreground very difficult. This necessitated the use of more advanced and time consuming filtering techniques such as the application of a garbage matte filter in order to completely remove the green background. More attention paid to good lighting, a uniform green background colour, and

clear contrast between foreground and background would simplify and hasten the video editing process.

Figure 27 also illustrates an uneven shoulder alignment, with the speaker's right shoulder being higher than the left. As the speaker was wearing a black shirt, the use of a black background in the final video helped to mask out the difference in shoulder alignment. The black background also helped to mask out the residue left behind when the green background was keyed out. A grey or transparent background colour in the final video would have been more appealing, but would only be possible with a consistent shoulder position and more attention paid to background and lighting as described above.

3.3 Video Editing Procedures

A number of video editing techniques for stabilising the video transition between word pairs were investigated.

Alpha channel masking

A video is essentially a series of picture frames with each picture being made up of a number of pixels. The 720p50 video format (Table 4) uses 32 bit colour depth. The first 24 bits (3 channels) are used to define the RGB colour of each pixel, while the last 8 bits, which is known as the alpha channel (Porter & Duff, 1984), defines the transparency. When two picture layers are combined, the alpha channel acts as a mask by defining how much of the underlying picture shows through the overlying picture.

Figure 28 - Alpha channel mask

In order to stabilise the transitions between word pair video frames, an alpha channel mask was applied to the eyes, nose and mouth region of the face. Adobe After Effects CS4 (abbreviated here as AAE) was used to apply the mask to each video frame. In this way, the essential facial features (eyes, nose, mouth) associated with speech showed through the mask, while the outline of the face (hair, ears, neck) remained still. Unfortunately, the jumps in facial position

between video frames against the static outline of the head created unnatural looking facial movements in the final video output.

Head stabilisation algorithms

An attempt was made to smooth out the jumps between word pairs using video image stabilisation algorithms. AAE contains a number of built-in algorithms for video image stabilisation. Options are available to stabilise position, rotation and scale. Each stabilisation algorithm functions in a similar way; a point or series of points on the video image are nominated and then the software adjusts the video frames in order to maintain the chosen point(s) in a fixed position, rotation or scale.

Figure 29 - Head position stabilisation algorithm

Figure 29 illustrates an attempt at head position stabilisation by fixing a point on each eyebrow. AAE adjusted each frame in the video in order to keep "Track Point 1" and "Track Point 2" in a fixed position in the final video output. Various combinations of tracking points were trialled including the eyebrows, eyes, ears, nose, mouth and neck. The results were similar in each case; the tracking points remained in a solidly fixed position while the rest of the face jumped up and down and from side to side in an unnatural looking manner.

It became clear that physical control of the head position was the most important factor in determining smooth transitions between word pair video frames. If the physical movements of the head position are too great, the image stabilisation algorithms will not be able to compensate for it. While the attempts at image stabilisation were not entirely successful, they did show promise for the fine tuning of the final video output. Further investigation of video image stabilisation techniques is recommended (see 3.8 Future Development).

3.4 Video Transition Analysis

With 400 word pairs that can be used in combination to form 100,000 unique sentences, it is inevitable that some sentences will sound more naturally spoken than others. A natural sounding sentence requires smooth transitions between adjacent word pairs. Some previous matrix sentence tests have excluded unnatural sounding sentences from their final test lists. The developers of the British version (Hall, 2006; Hewitt, 2007) randomly generated lists of test sentences, which were then evaluated to ensure they sounded naturally spoken and did not contain clicks or other audible unwanted inclusions. Hall (2006) found that their female speaker had difficulty pronouncing the word pair "cheap chairs", and this error was carried through to the final sentences. Hall (2006) therefore removed any sentences with "cheap chairs" and generated new sentences in their place. The developers of the Spanish matrix sentence test (Hochmuth et al., 2012) evaluated different concatenation overlap durations between successive word pairs. Hochmuth et al. (2012) found that some sentences had a better sound quality using an overlap of 15 ms, while for other sentences, fewer artefacts were noticed when no overlap was used. Only those sentences with the least perceivable artefacts were used for the further development of the test.

Adding a video component to the UC Auditory-visual Matrix Sentence Test creates more complexity as both the audio and video components need to sound and look naturally spoken. Unlike previous matrix sentence tests, no ramping or overlapping of the sound files was used. While these methods were trialled, a more natural sounding audio transition was achieved through careful selection of the word pair cutting points. Furthermore, the absence of ramps and overlaps allowed the audio and video components to remain synchronized. Very few sentences were found in which the video component looked natural, while the audio component sounded unnatural. However, sentences containing the word "red" were found to be one such example. The "red" waveform does not contain a sustained interval of near-zero amplitude and must therefore be cut in the middle of the waveform. When combined with other word pairs, the mismatch in

waveform amplitude can create an audible click. Sentences in which the audio component sounds natural while the video component looks unnatural were much more common.

A method for eliminating the worst video transitions by comparing the difference in RGB colour is described in section 2.7. "Nine-good ships" was found to be the best video transition. "Peter has nine good ships" was part of the 100 originally recorded sentences and therefore "Nine-good ships" is naturally spoken and not made up of word pairs from separate sentences. The worst transitions were found to be those containing the word pair "eight-big". Appendix A shows the "eight" waveform in the originally recorded "Oscar bought eight big mugs" has the greatest amplitude of all of the "eight" waveforms. It was the fifth of one hundred sentences and perhaps the burst of energy coincided with the speaker realising she was at the start of the sentence list for the third and final time. While the reason for the extra effort is open for debate, the greater waveform amplitude and open mouth position created the largest difference between adjoining word pairs, and therefore the worst transition.

Two sets of video transition analysis were made; the first including the mouth position and the second excluding. The mouth region contains the most movement and therefore by excluding it from the analysis the absolute difference between video frames is reduced (Figure 24). When the mouth region is excluded the difference between video frames can be mainly attributed to changes in head position.

As more word pairs are excluded, the number of available sentences is reduced. Although this function has been modelled (Figure 25), care needs to be taken with the interpretation. One cannot simply exclude 50% of the word pairs and think the remaining 10,000 sentences will be more than enough sentence material. Although there may be many unique sentences remaining, they will be very repetitive as the same word pairs are used over and over again. Also, by excluding

word pairs the phonemic distribution of the matrix becomes increasingly unbalanced. Straight subtraction of RGB colour channels may not be the best way to evaluate the smoothness of the video transitions. The comparison of motion vectors between video frames may be more appropriate and further investigation into video image analysis techniques is recommended.

3.5 Clinical Applications

Matrix sentence tests in the auditory-alone modality have found clinical application where repeated measures of speech audiometry on the same subject are required. The matrix sentences are unlikely to be memorised in contrast to other sentences test such as everyday sentences, Plomp sentences and HINT sentences. This makes matrix sentences very useful for hearing aid evaluations. The simple 50 word structure also makes matrix sentences useful for evaluating cochlear implants users, and testing children.

In addition to the applications described above, the UC Auditory-visual Matrix Sentence Test will allow the evaluation of lip-reading and auditory-visual integration abilities. As described in section 1.9, the auditory-visual integration abilities of the hearing impaired vary greatly between individuals. The measurement of these abilities will help to provide counselling on realistic expectations as to the likely benefits of sensory aids, and a better understanding of the listening environments where such aids are likely to be effective. Ultimately, the assessment of auditory-visual integration abilities can be used to develop individually based aural rehabilitation strategies.

3.6 Research Applications

The clinical uses of the UC Auditory-visual Matrix Sentence Test also extend themselves into the research environment, where studies involving repeated measures are more common. The practically endless supply of sentences makes

the UC Auditory-visual Matrix Sentence Test a powerful tool for researcher's needing repeat measures of speech audiometry.

While conventional speech-in-noise tests have used multi-talker babble and speech spectrum masking noise, the body of evidence in the literature (see 1.4 Masking Noise) suggests that amplitude modulated and/or noise with spectral gaps may provide better separation of normal hearing and hearing impaired listeners. Four different masking noise options will be made available in the UC Auditory-visual Matrix Sentence Test:

1. Continuous octave band noise
2. Octave band noise with spectral gaps
3. Amplitude modulated noise
4. Amplitude modulated noise with spectral gaps

The specific characteristics and implementation of the masking noise options will require further investigation. However, once implemented, it will allow research to be conducted in the auditory-alone, visual-alone and auditory-visual modalities in order to find the masking parameters that provide the best sensitivity and specificity for separating normal hearing and hearing impaired groups of listeners.

3.7 Limitations

Matrix sentence tests in the auditory-alone modality tend to lack the real world context that is provided by everyday sentences, Plomp sentences and HINT sentences. In contrast to monosyllabic speech tests, matrix sentence tests require good working memory to store and recall the 5 word sentences. The same limitations apply to the UC Auditory-visual Matrix Sentence Test. In addition, the visual stimuli is more susceptible to looking unnatural, than the auditory stimuli is to sounding unnatural. The human visual system is very sensitive to the kind of rapid movements created when there is a mismatch between video frames (Cropper & Derrington, 1996).

3.8 Future Development

The future development of the UC Auditory-visual Matrix Sentence Test needs to progress along three pathways. Firstly, there needs to be further refinement of the procedures for stabilising the head position, which is essential in order that the sentences in the final video output look naturally spoken. Secondly, the core software application which mixes the audio and video files together to form sentences needs to be integrated with the University of Canterbury Adaptive Speech (UCAST) test platform. The UCAST already contains an adaptive procedure for detecting speech reception threshold, the ability to add masking noise, and scoring procedures. Thirdly, the auditory and visual speech stimuli need to be normalised.

Stabilisation of head position

The speaker's head needs to be held in a constant position throughout the entire sentence recording session. If this can be achieved, the transition between word pair video frames will appear smooth and naturally spoken in the final video output. The head support system illustrated in Figure 26 was very basic in order to prove the concept. More advanced head support systems need to be investigated. The ideal head support system would provide support behind the neck and shoulder region such that the speaker is not able to slump down as the recording session proceeds. The ideal head support system would also prevent the speaker from straining to maintain an upright body position, which would help to maintain a more constant facial expression.

One possibility is to use a head alignment clamp that has been developed previously for eye tracking studies (Figure 30).

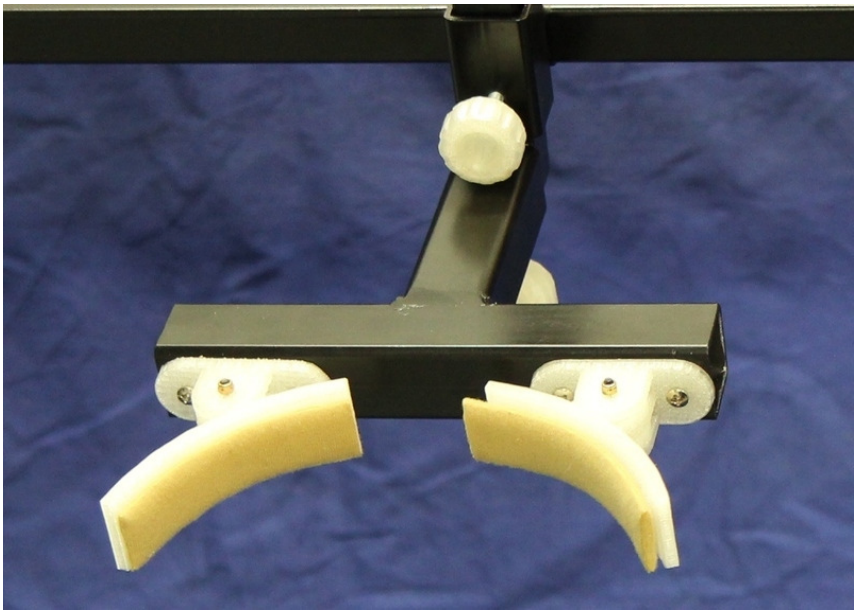


Figure 30 - Head alignment clamp (Swosho, 2012)

The advantage of a support behind the head is that it can be hidden from view or easily edited out with a green screen. However, the disadvantage is that it may not provide enough support to prevent the subtle head movements that can ruin a video recording session.

Another possibility is to use a head brace that has been developed previously for the treatment of neck and spinal injuries.

Figure 31 - Halo head brace (Bremer Medical Incorp, Jacksonville, FL, USA)

The advantage of a halo head brace is that the head is supported at all angles, which should result in a very stable head position. The disadvantage is the effort required to edit the brace out of the final video output. However, the alpha channel masking techniques described in section 3.3 would be suitable in this case. An image of the speaker without the head brace could be used to mask out the brace during the post recording editing process.

While the bulk of the head movement needs to be controlled with a physical support, there is the potential for fine tuning of the final video output with image stabilisation algorithms. Experiments using algorithms that attempt to stabilise one or two reference points in the video have not been successful (see 3.3 Head stabilisation algorithms). Algorithms that compare entire video frames and attempt to find the best average between them should be investigated. A lot of computing power is required for editing video footage, especially in high

definition at frame rates of 50 fps. It is recommended that the highest specification computer processor and graphic card that resources will allow be used for the editing of future versions. A 64 bit Windows operating system is also recommended as the latest versions of video editing software suites such as Adobe no longer support the 32 bit version of Windows.

Once the techniques for head stabilisation have been finalised, a subjective rating exercise to assess the "naturalness" of the UC Auditory-visual matrix sentences will be undertaken with a group of 20 - 30 normal hearing listeners. Listeners will be presented with 50 actual sentences (i.e. original sentences voiced by the speaker), and 50 synthesised sentences, all in randomised order. The listeners will subjectively rate the sentences on a scale of 1 – 10 (1 = very unnatural, 10 = very natural). The procedure will be repeated in auditory-alone, visual-alone, and auditory-visual modes. The results will be compared against the objective measures of video image stability described in section 2.7.

Integration with University of Canterbury Adaptive Speech Test

The University of Canterbury Adaptive Speech Test (UCAST; O'Beirne, McGaffin, & Rickard, 2012) is based on the Monosyllabic Adaptive Speech Test of Mackie and Dermody (1986). The UCAST was developed as an adaptive, low-pass filtered speech test whereby the user selects one of four choices from a touch screen.

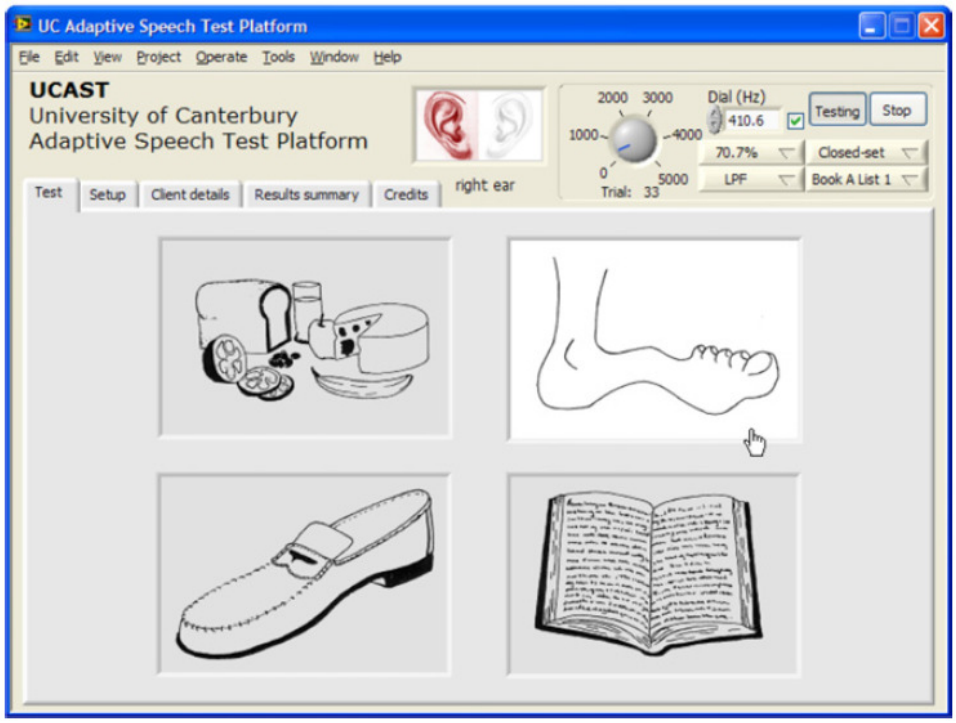


Figure 32 - University of Canterbury Adaptive Speech Test user interface

The UCAST was developed in the LabVIEW environment and includes programming for the addition of masking noise, adaptive threshold seeking procedures, touch screen functionality and scoring. The UCAST is being developed into suite of audiological tests which will include the NZHINT (Hope, 2010) and New Zealand Digit Triplet Test (NZDTT; King, 2011). The NZDTT is a hearing screening tool that uses spoken numbers presented in background noise to estimate speech recognition thresholds. The UC Auditory-visual Matrix Sentence Test will also be integrated with, and include the functionality of, the UCAST platform.

Normalisation

The normalisation of the UC Auditory-visual matrix sentence stimuli presents specific challenges because of the way the sentence material is edited into fragments. Normalisation is aimed at ensuring that each audio or video fragment is equally difficult, and this in turn ensures that sentences made from these fragments are also equally difficult. Because scoring a response correct or incorrect is done at the word level, the score must be mapped onto the audio or video files that contain that that particular material. The sentence "Amy bought two big bikes" is formed from four files, named amy_bought, bought_two, two_big, and big_bikes. Similarly, "William wins those small toys" is made from william_wins, wins_those, those_small, and small_toys. However, due to the complicated editing process, the way the written words correspond to the sounds in the files is quite different in each case, as shown below:

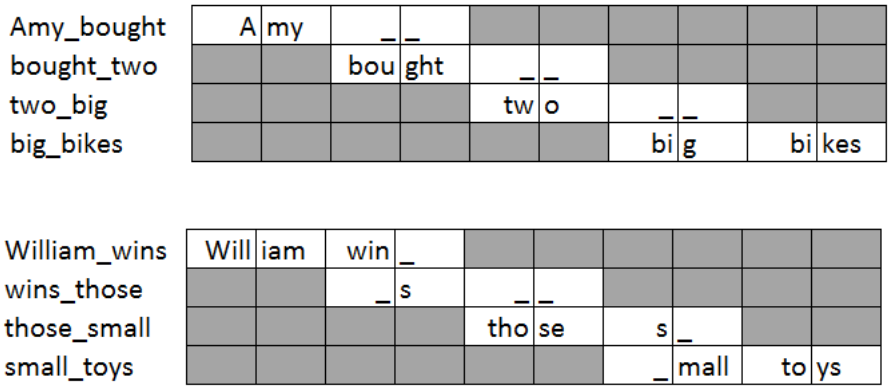


Figure 33 - Word pair vs sound file contents

The scoring procedures for the UCAST have been coded in the LabVIEW environment based on binary number operations. In order to integrate the UC Auditory-visual Matrix Sentence Test with the existing binary functions of the UCAST, the following scoring procedure is proposed:

Example 1: "Amy bought two big bikes"

Sample

Amy_bought	A	my	_	_						
bought_two			bou	ght	_	_				
two_big					tw	o	_	_		
big_bikes							bi	g	bi	kes

Amy_bought	1A	1A	0B	0B							Sum:	2A	0B
bought_two			1A	1A	0B	0B					Sum:	2A	0B
two_big					1A	1A	0B	0B			Sum:	2A	0B
big_bikes							1A	1A	1B	1B	Sum:	2A	2B

Sentence	Amy	bought	two	big	bikes
Selected	Amy	bought	those	big	shirts
Correct?	1	1	0	1	0

Amy_bought	1A	1A	0B	0B							Sum:	2A	0B
bought_two			1A	1A	0B	0B					Sum:	2A	0B
two_big					0A	0A	0B	0B			Sum:	0A	0B
big_bikes							1A	1A	0B	0B	Sum:	2A	0B

Score

Amy_bought	2A/2A =	100% A (Amy)	0B/0B =	0% B (bought)
bought_two	2A/2A =	100% A (bought)	0B/0B =	0% B (two)
two_big	0A/2A =	0% A (two)	0B/0B =	0% B (big)
big_bikes	2A/2A =	100% A (big)	0B/2B =	0% B (bikes)

Figure 34 - Scoring of "Amy bought two big bikes"

The word pairs are divided into parts A and B in order to track their location with each sound file. Figure 34 shows that Amy_bought (Part A_part B) is represented by "A" + "my" = 1A + 1A = 2A, while "bought" is represented by "bou" + "ght" = 0B + 0B = 0B. The zero indicates that the "bought" portion of the sound file is not actually contained within Amy_bought. Instead, the "bought" portion is located in part A of the bought_two sound file. When a user makes an incorrect response, e.g. "Amy bought those big shirts" instead of "Amy bought two big bikes", the scores for the incorrect words ("those" and "shirts") are set to zero within the binary matrix. The user's response is then compared to the actual sentence presented in order to calculate the score e.g.

["shirts" (incorrect) = 0B] / ["bikes" (correct) = 2B] = 0% for part B of big_bikes

Example 2: "William wins those small toys"

Sample

William_wins	Will	iam	win	_						
wins_those			_	s	_	_				
those_small					tho	se	s	_		
small_toys							_	mall	to	ys

William_wins	1A	1A	1B	0B							Sum:	2A	1B
wins_those			0A	1A	0B	0B					Sum:	1A	0B
those_small					1A	1A	1B	0B			Sum:	2A	1B
small_toys							0A	1A	1B	1B	Sum:	1A	2B

Sentence	William	wins	those	small	toys
Selected	William	wins	four	small	shoes
Correct?	1	1	0	1	0

William_wins	1A	1A	1B	0B							Sum:	2A	1B
wins_those			0A	1A	0B	0B					Sum:	1A	0B
those_small					0A	0A	1B	0B			Sum:	0A	1B
small_toys							0A	1A	0B	0B	Sum:	1A	0B

Score

William_wins	2A/2A =	100% A (William)	1B/1B =	100% B (wins)
wins_those	1A/1A =	100% A (wins)	0B/0B =	0% B (those)
those_small	0A/2A =	0% A (those)	1B/1B =	100% B (small)
small_toys	1A/1A =	100% A (small)	0B/2B =	0% B (toys)

Figure 35 - Scoring of "William wins those small toys"

Figure 35 shows that the sound "wins" is divided across the two files William_wins and wins_those. The "win" portion is contained in part B of William_wins while the "s" portion is contained in part A of wins_those. If the user selects "wins" as a response, the score is represented by 1B in the William_wins file and by 1A in the wins_those file. Any scores that create a division by zero error (e.g. 0B / 0B those) are reassigned to 0%, which indicates an incorrect response.

A pilot study will be undertaken in order to determine the approximate signal-to-noise ratio (SNR) corresponding to 30, 50 and 70% intelligibility of the matrix sentences. The resulting three fixed SNR values will be tested with a group of normal hearing listeners to find the word specific speech intelligibility functions (Figure 36).

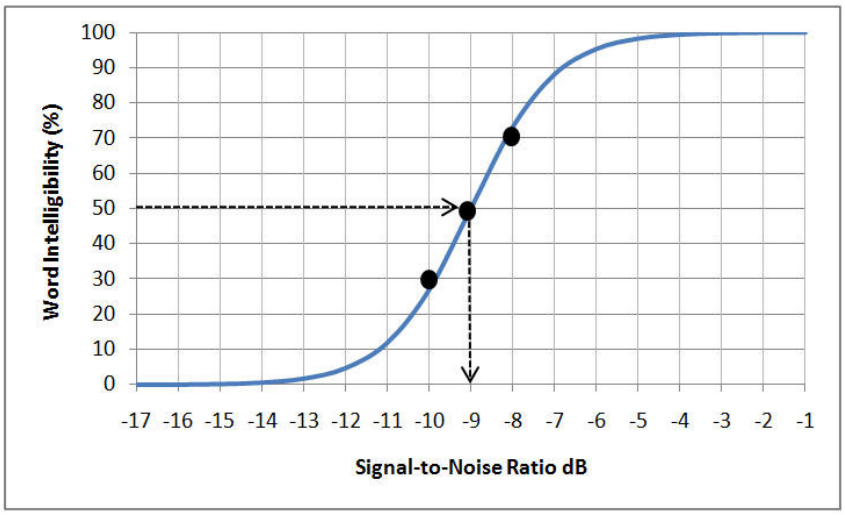


Figure 36 - Word specific intelligibility function

Each participant will be presented with 20 practice sentences in order to become familiar with the testing procedure and user interface. The participants will then be presented with 100 randomly generated sentences containing all 400 word pair combinations. A spreadsheet macro has been prepared to randomly generate lists of 100 sentences while ensuring all 400 word pairs are used once. Masking noise will be presented at a constant level of 65 dB SPL while the level of the sentence presentation is varied in order to achieve a SNR corresponding to 30, 50 or 70% intelligibility. Participants' responses will be scored using the procedure described above and stored by the user interface for analysis. Ten data points will be recorded for each SNR for each word pair combination. A total of 12,000 data points will be recorded (100 sentences x 4 word pairs per sentence x 3 SNR). The average SNR corresponding to 50% intelligibility of each word pair combination will be compared to the average SNR corresponding to 50% intelligibility of the sentences (Figure 37).

	William	wins	those	small	toys	Sentence Intelligibility
Participant 1	100%	100%	0%	0%	100%	60%
Participant 2	100%	0%	100%	100%	0%	60%
Participant 3	0%	100%	0%	0%	100%	40%
Participant 4	100%	100%	100%	0%	0%	60%
Participant 5	0%	0%	100%	100%	100%	60%
Participant 6	100%	0%	100%	0%	0%	40%
Participant 7	100%	0%	100%	0%	0%	40%
Participant 8	0%	0%	0%	100%	100%	40%
Participant 9	0%	100%	0%	100%	100%	60%
Participant 10	100%	0%	100%	0%	100%	60%
Word Intelligibility	60%	40%	60%	40%	60%	52%

Figure 37 - Participants' word intelligibility vs sentence intelligibility (hypothetical example)

The levels of each word pair will be adjusted up or down in order to match the average intelligibility of the sentences. From the example given in Figure 37, the level of "William" needs to be decreased while the level of "wins" needs to be increased in order to match the sentence intelligibility. Level adjustments will be limited to ±4 dB as previous studies (Wagener et al., 2003; Ozimek et al., 2010) have found this to be the maximum allowable level adjustment that can be made without causing the sentences to sound unnatural. Equalising the intelligibility of each word in the matrix should result in sentences of equal intelligibility. Another group of normal hearing listeners will be required to confirm this hypothesis. Studies comparing performance in auditory-alone, visual-alone and auditory-visual modes should also be undertaken.

4 Conclusions

- The UC Auditory-visual Matrix Sentence Test holds promise as a multi-modal test of speech perception.

- Presentation in auditory-alone, visual-alone, and auditory-visual modes allows assessment of auditory-visual integration abilities.

- A stable head position and consistent facial expression are essential for a natural looking synthesized sentence.

- More research is required into both physical and mathematical methods for stabilising head position.

Appendix 1 – New Zealand Matrix Sentence Recording List

Amy	bought	two	big	bikes
David	bought	three	big	books
Hannah	bought	four	big	coats
Kathy	bought	six	big	hats
Oscar	bought	eight	big	mugs
Peter	bought	nine	big	ships
Rachel	bought	ten	big	shirts
Sophie	bought	twelve	big	shoes
Thomas	bought	some	big	spoons
William	bought	those	big	toys
Amy	gives	two	cheap	bikes
David	gives	three	cheap	books
Hannah	gives	four	cheap	coats
Kathy	gives	six	cheap	hats
Oscar	gives	eight	cheap	mugs
Peter	gives	nine	cheap	ships
Rachel	gives	ten	cheap	shirts
Sophie	gives	twelve	cheap	shoes
Thomas	gives	some	cheap	spoons
William	gives	those	cheap	toys
Amy	got	two	dark	bikes
David	got	three	dark	books
Hannah	got	four	dark	coats
Kathy	got	six	dark	hats
Oscar	got	eight	dark	mugs
Peter	got	nine	dark	ships
Rachel	got	ten	dark	shirts
Sophie	got	twelve	dark	shoes
Thomas	got	some	dark	spoons
William	got	those	dark	toys
Amy	has	two	good	bikes
David	has	three	good	books
Hannah	has	four	good	coats
Kathy	has	six	good	hats
Oscar	has	eight	good	mugs
Peter	has	nine	good	ships
Rachel	has	ten	good	shirts
Sophie	has	twelve	good	shoes
Thomas	has	some	good	spoons
William	has	those	good	toys
Amy	kept	two	green	bikes
David	kept	three	green	books
Hannah	kept	four	green	coats
Kathy	kept	six	green	hats
Oscar	kept	eight	green	mugs
Peter	kept	nine	green	ships
Rachel	kept	ten	green	shirts
Sophie	kept	twelve	green	shoes
Thomas	kept	some	green	spoons
William	kept	those	green	toys

Amy	likes	two	large	bikes
David	likes	three	large	books
Hannah	likes	four	large	coats
Kathy	likes	six	large	hats
Oscar	likes	eight	large	mugs
Peter	likes	nine	large	ships
Rachel	likes	ten	large	shirts
Sophie	likes	twelve	large	shoes
Thomas	likes	some	large	spoons
William	likes	those	large	toys
Amy	sees	two	new	bikes
David	sees	three	new	books
Hannah	sees	four	new	coats
Kathy	sees	six	new	hats
Oscar	sees	eight	new	mugs
Peter	sees	nine	new	ships
Rachel	sees	ten	new	shirts
Sophie	sees	twelve	new	shoes
Thomas	sees	some	new	spoons
William	sees	those	new	toys
Amy	sold	two	old	bikes
David	sold	three	old	books
Hannah	sold	four	old	coats
Kathy	sold	six	old	hats
Oscar	sold	eight	old	mugs
Peter	sold	nine	old	ships
Rachel	sold	ten	old	shirts
Sophie	sold	twelve	old	shoes
Thomas	sold	some	old	spoons
William	sold	those	old	toys
Amy	wants	two	red	bikes
David	wants	three	red	books
Hannah	wants	four	red	coats
Kathy	wants	six	red	hats
Oscar	wants	eight	red	mugs
Peter	wants	nine	red	ships
Rachel	wants	ten	red	shirts
Sophie	wants	twelve	red	shoes
Thomas	wants	some	red	spoons
William	wants	those	red	toys
Amy	wins	two	small	bikes
David	wins	three	small	books
Hannah	wins	four	small	coats
Kathy	wins	six	small	hats
Oscar	wins	eight	small	mugs
Peter	wins	nine	small	ships
Rachel	wins	ten	small	shirts
Sophie	wins	twelve	small	shoes
Thomas	wins	some	small	spoons
William	wins	those	small	toys

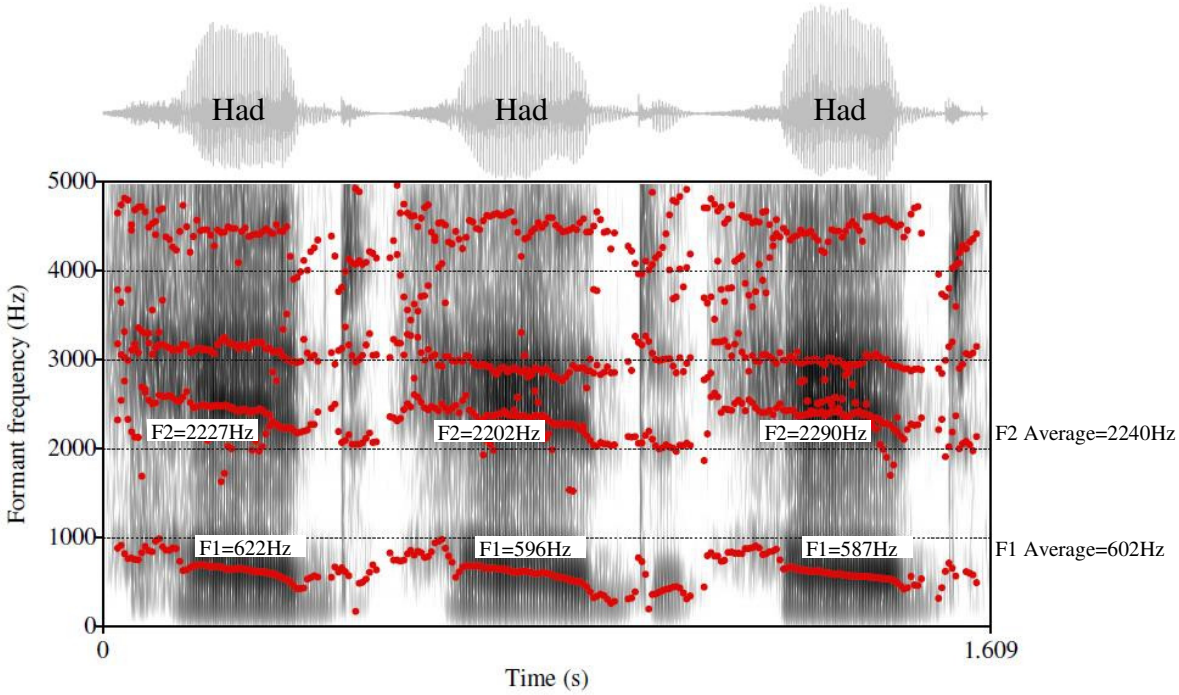
Appendix 2 – Phonemic Distribution Analysis

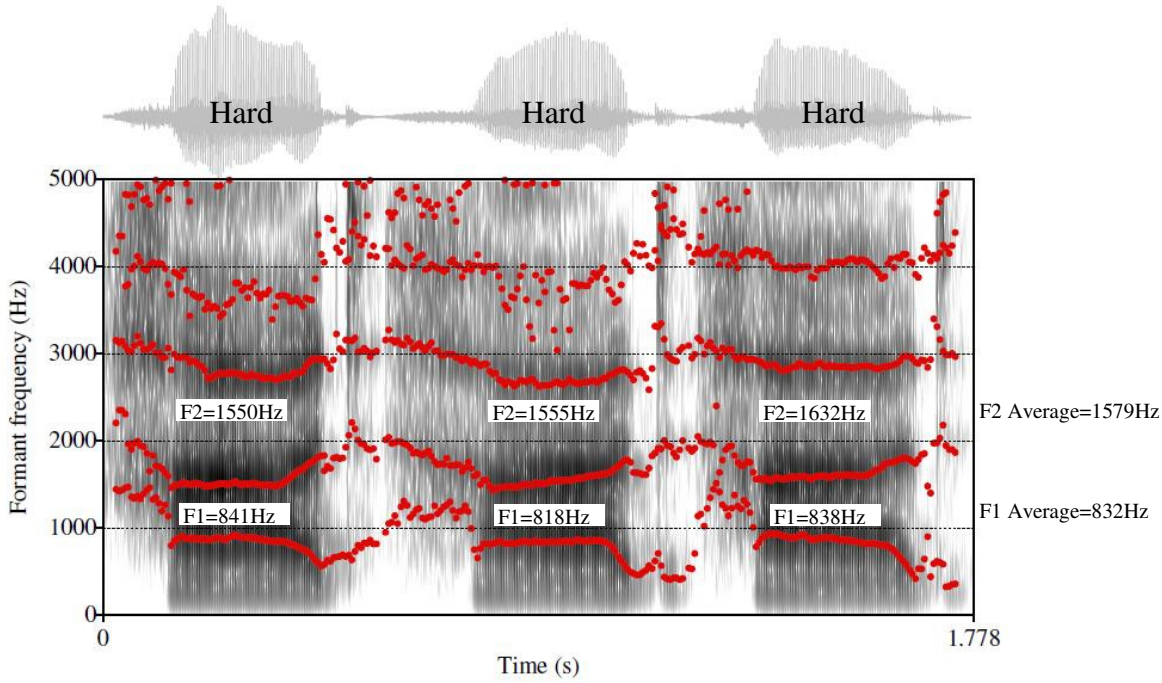
	p	t	k	b	d	g	f	th	v	dth	n	m	ng	l	r	tʃ	dj	s	z	ʃ	w	h	y
Peter	1	1																					
Kathy			1					1															
Sophie							1											1					
David					2				1														
Rachel														1	1	1							
Amy												1											
William												1		1							1		1
Thomas		1										1						1					
Hannah											1											1	
Oscar			1															1					
got		1				1																	
sees																		1	1				
bought		1		1																			
gives						1			1										1				
sold					1									1				1					
likes			1											1				1					
has																			1			1	
kept	1	1	1																				
wins											1								1		1		
wants		1									1							1			1		
three								1							1								
those										1									1				
nine											2												
eight		1																					
four							1																
six			1															2					
two		1																					
ten		1									1												
twelve		1							1					1							1		
some												1						1					
large														1			1						
small												1		1				1					
old					1									1									
dark			1		1																		
good					1	1																	
green						1					1				1								
cheap	1															1							
new											1												1
red					1										1								
big				1		1																	
books			1	1														1					
coats		1	1															1					
shoes																			1	1			
toys		1																	1				
spoons	1										1							1	1				
mugs						1						1							1				
ships	1																	1		1			
hats		1																1				1	
shirts		1																1		1			
bikes			1	1														1					
Total	5	14	9	4	7	6	2	2	3	1	9	6	0	8	4	2	1	18	9	3	4	3	2
%age	2.7%	7.6%	4.9%	2.2%	3.8%	3.3%	1.1%	1.1%	1.6%	0.5%	4.9%	3.3%	0.0%	4.3%	2.2%	1.1%	0.5%	9.8%	4.9%	1.6%	2.2%	1.6%	1.1%
NZHINT	2.1%	7.6%	4.1%	1.8%	4.8%	1.1%	1.8%	0.4%	1.2%	6.3%	4.9%	2.7%	0.9%	4.1%	2.5%	0.6%	0.3%	4.2%	2.8%	1.6%	3.3%	2.5%	0.3%

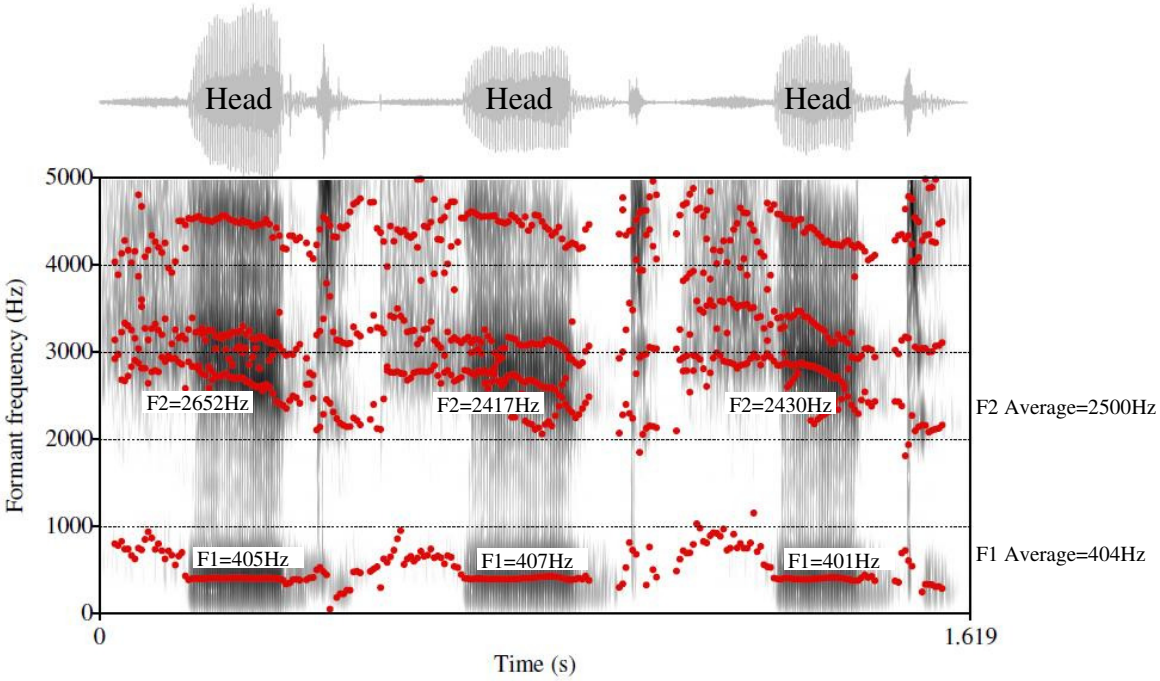
Phonemic Distribution Analysis - Continued

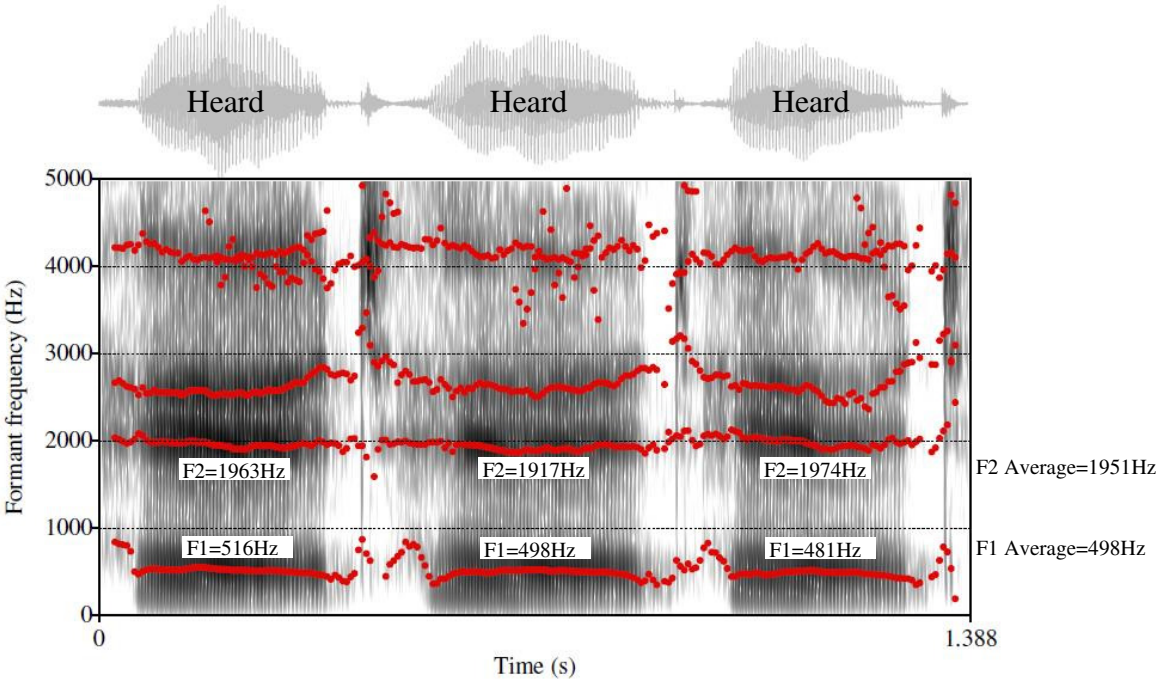
	KIT	FLEECE	DRESS	TRAP	LOT	FORCE	START	NURSE	GOOSE	STRUT	FOOT	FACE	PRICE	GOAT	CHOICE	MOUTH	NEAR	GOLD	SCHWA
Peter		1																	1
Kathy		1		1															
Sophie		1												1					
David												1							1
Rachel												1							1
Amy		1		1								1							1
William	1																		1
Thomas					1														1
Hannah				1															1
Oscar					1														1
got					1														
sees		1																	
bought						1													
gives	1																		
sold																		1	
likes													1						
has				1															
kept			1																
wins	1																		
wants					1														
three		1																	
those														1					
nine													1						
eight												1							
four						1													
six	1																		
two									1										
ten			1																
twelve			1																
some										1									
large							1												
small						1													
old																		1	
dark							1												
good											1								
green		1																	
cheap		1																	
new									1										
red			1																
big	1																		
books											1								
coats														1					
shoes									1										
toys															1				
spoons									1										
mugs										1									
ships	1																		
hats				1															
shirts								1											
bikes													1						
Total	6	8	4	5	4	3	2	1	4	2	2	4	3	3	1	0	0	2	8
%age	3.3%	4.3%	2.2%	2.7%	2.2%	1.6%	1.1%	0.5%	2.2%	1.1%	1.1%	2.2%	1.6%	1.6%	0.5%	0.0%	0.0%	1.1%	4.3%
NZHINT	1.9%	4.3%	1.7%	1.9%	2.3%	1.3%	0.8%	0.8%	1.0%	1.7%	0.6%	2.8%	1.4%	1.0%	0.2%	0.9%	0.6%	0.4%	12.7%

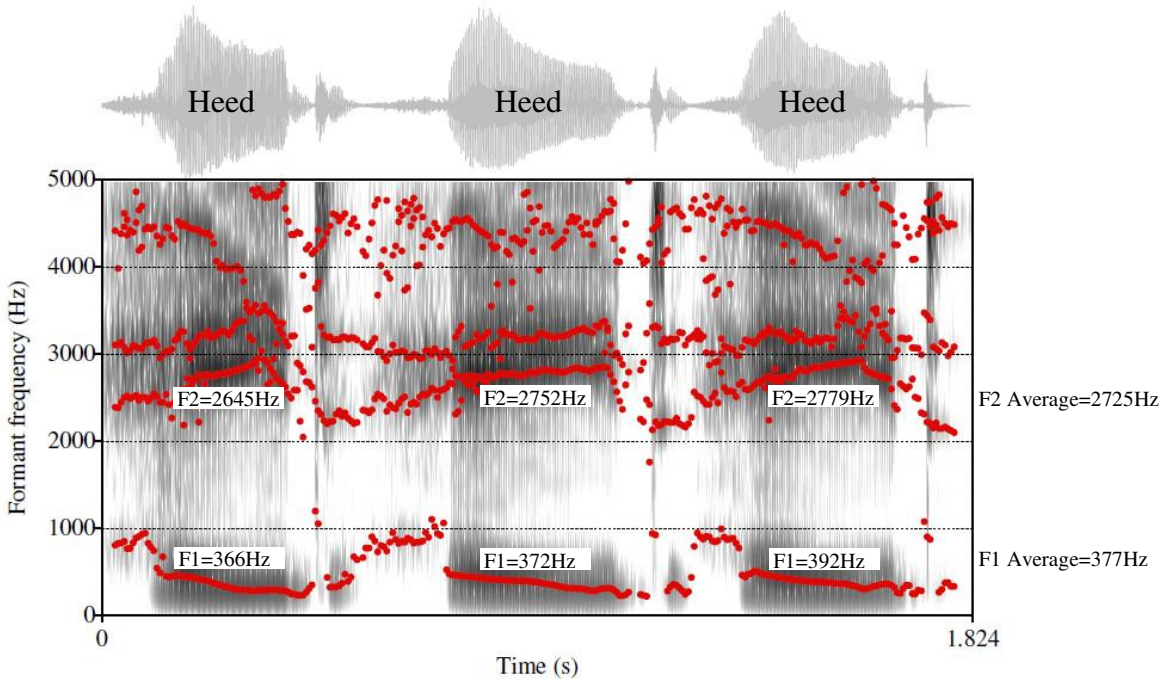
Appendix 3 – Vowel Formant Frequency Analysis

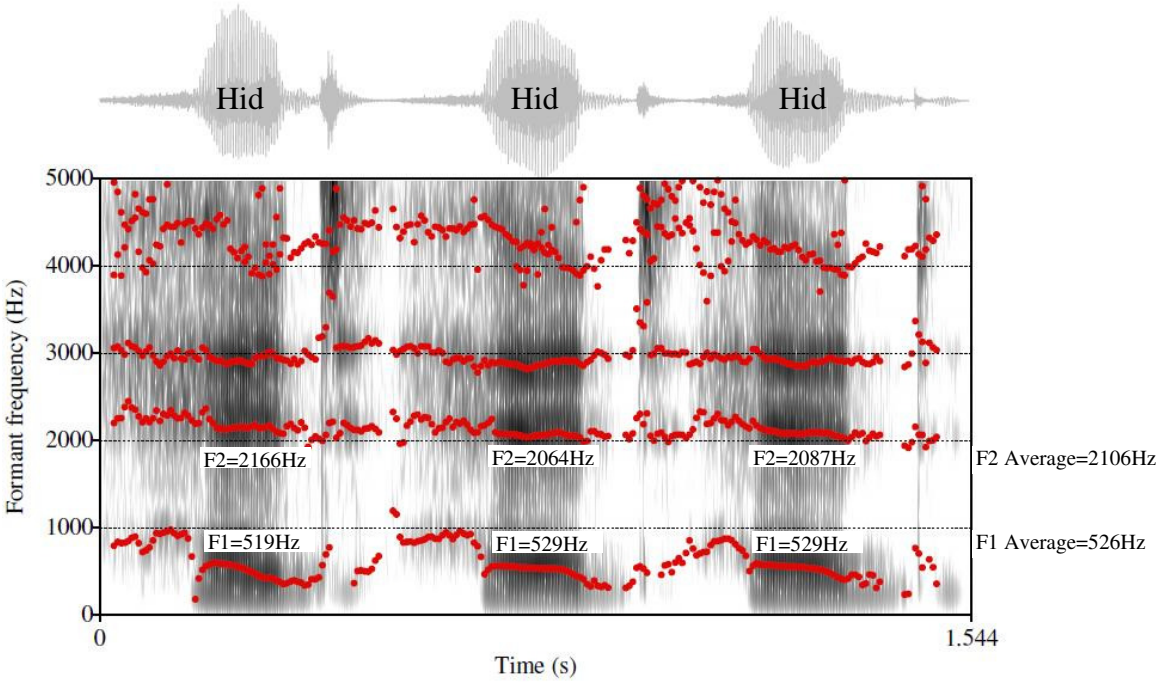

Had
Had
Had
5000
4000
3000
2000
1000
0
Formant frequency (Hz)
F2=2227Hz
F2=2202Hz
F2=2290Hz
F2 Average=2240Hz
F1=622Hz
F1=596Hz
F1=587Hz
F1 Average=602Hz
0
1.609
Time (s)

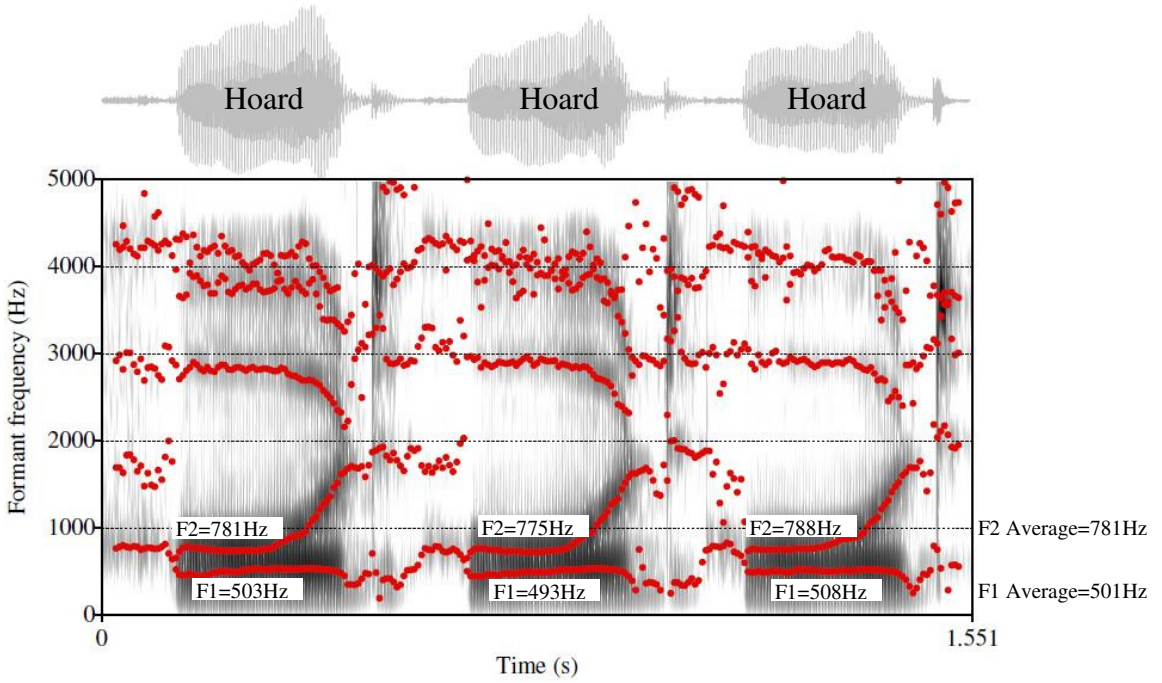

Hard
Hard
Hard
5000
4000
3000
2000
1000
0
Formant frequency (Hz)
F2=1550Hz
F2=1555Hz
F2=1632Hz
F2 Average=1579Hz
F1=841Hz
F1=818Hz
F1=838Hz
F1 Average=832Hz
0
1.778
Time (s)

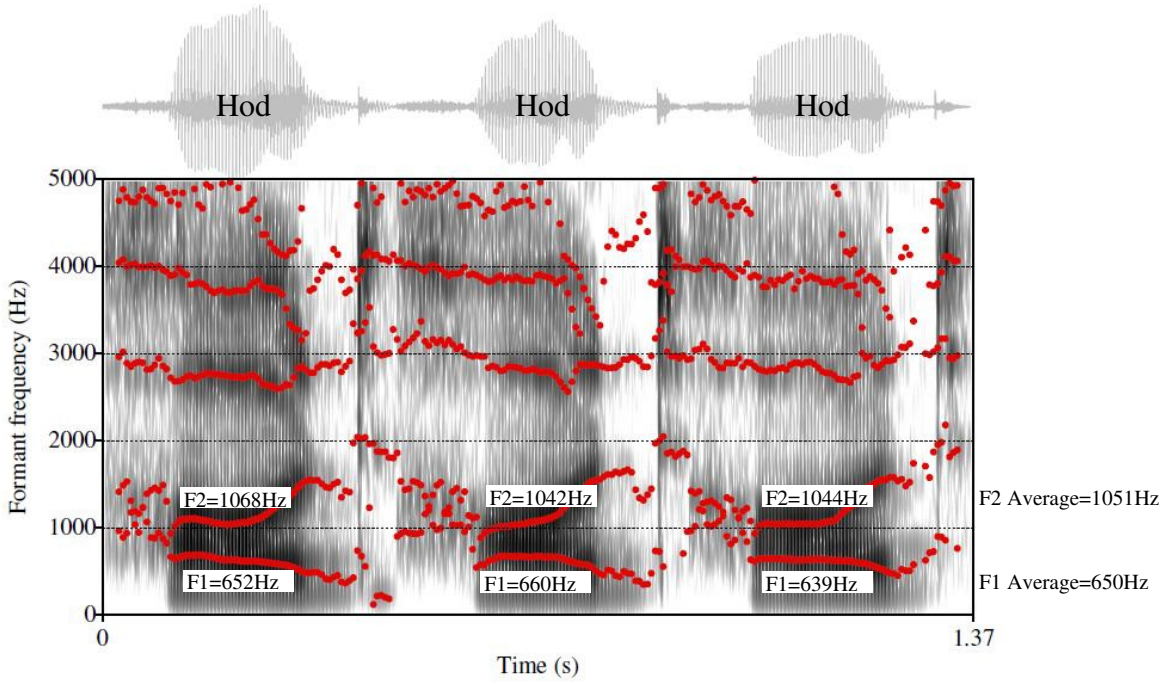

Head
Head
Head
5000
4000
3000
2000
1000
0
Formant frequency (Hz)
F2=2652Hz
F2=2417Hz
F2=2430Hz
F2 Average=2500Hz
F1=405Hz
F1=407Hz
F1=401Hz
F1 Average=404Hz
0
1.619
Time (s)

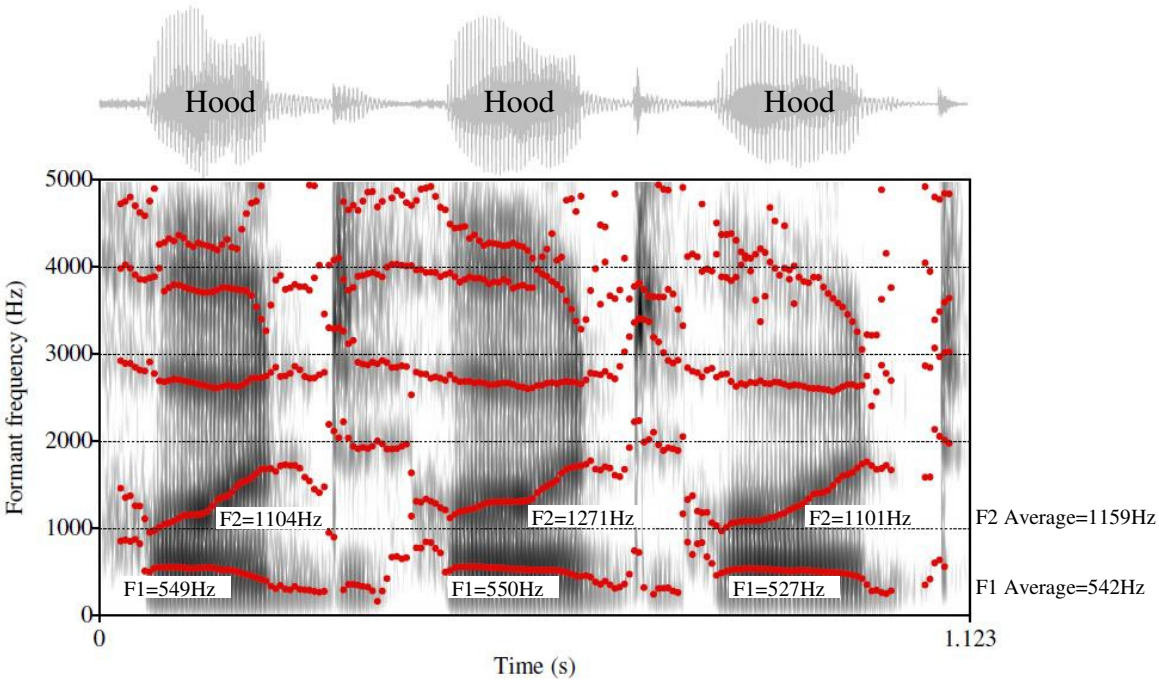

Heard
Heard
Heard
5000
4000
3000
2000
1000
0
Formant frequency (Hz)
F2=1963Hz
F2=1917Hz
F2=1974Hz
F2 Average=1951Hz
F1=516Hz
F1=498Hz
F1=481Hz
F1 Average=498Hz
0
1.388
Time (s)

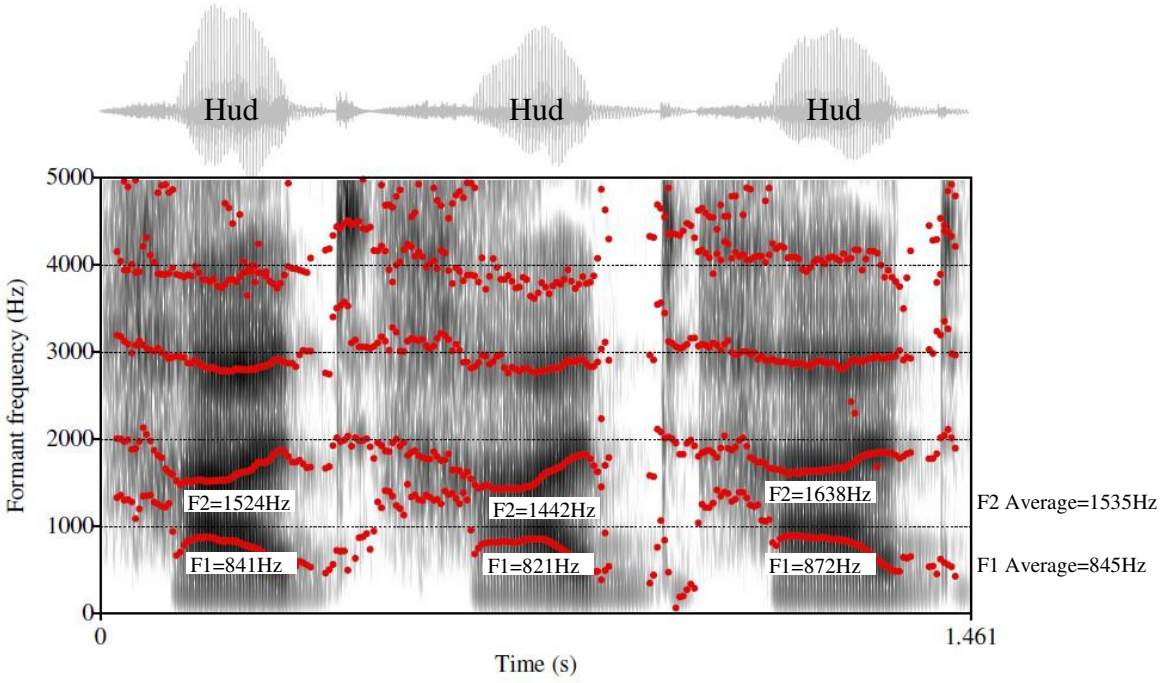

Heed
Heed
Heed
5000
4000
3000
2000
1000
0
Formant frequency (Hz)
F2=2645Hz
F2=2752Hz
F2=2779Hz
F2 Average=2725Hz
F1=366Hz
F1=372Hz
F1=392Hz
F1 Average=377Hz
0
1.824
Time (s)

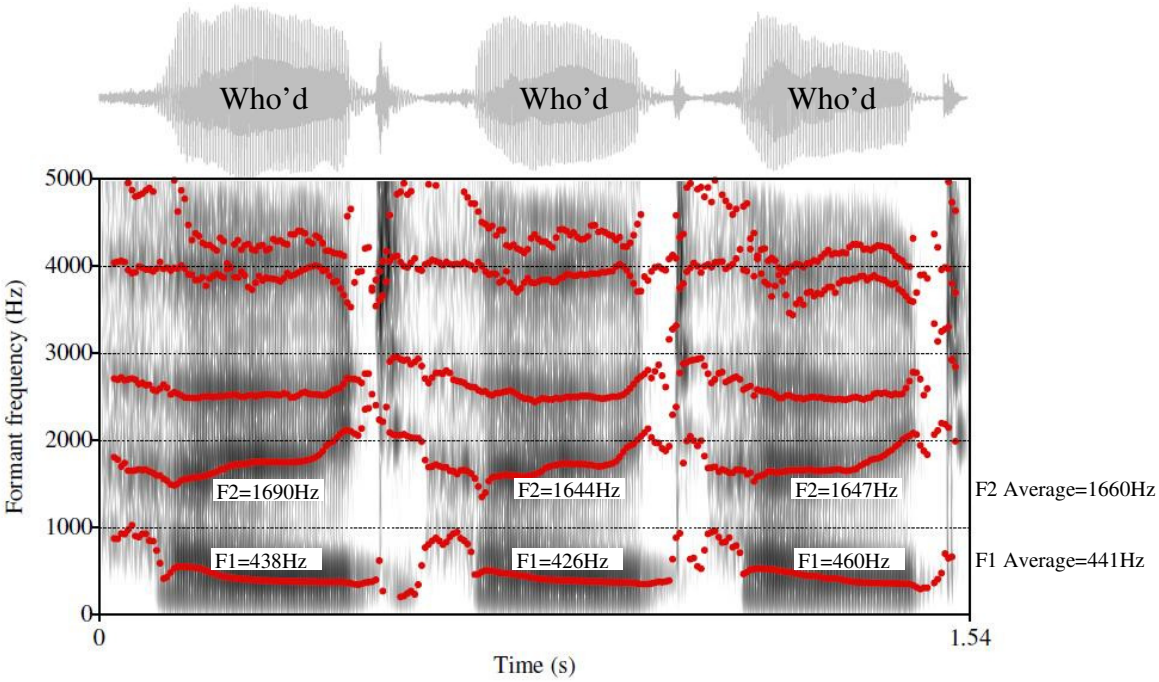

Hid
Hid
Hid
5000
4000
3000
2000
1000
0
Formant frequency (Hz)
F2=2166Hz
F2=2064Hz
F2=2087Hz
F2 Average=2106Hz
F1=519Hz
F1=529Hz
F1=529Hz
F1 Average=526Hz
0
1.544
Time (s)


Hoard
Hoard
Hoard
5000
4000
3000
2000
1000
0
Formant frequency (Hz)
F2=781Hz
F2=775Hz
F2=788Hz
F2 Average=781Hz
F1=503Hz
F1=493Hz
F1=508Hz
F1 Average=501Hz
0
1.551
Time (s)

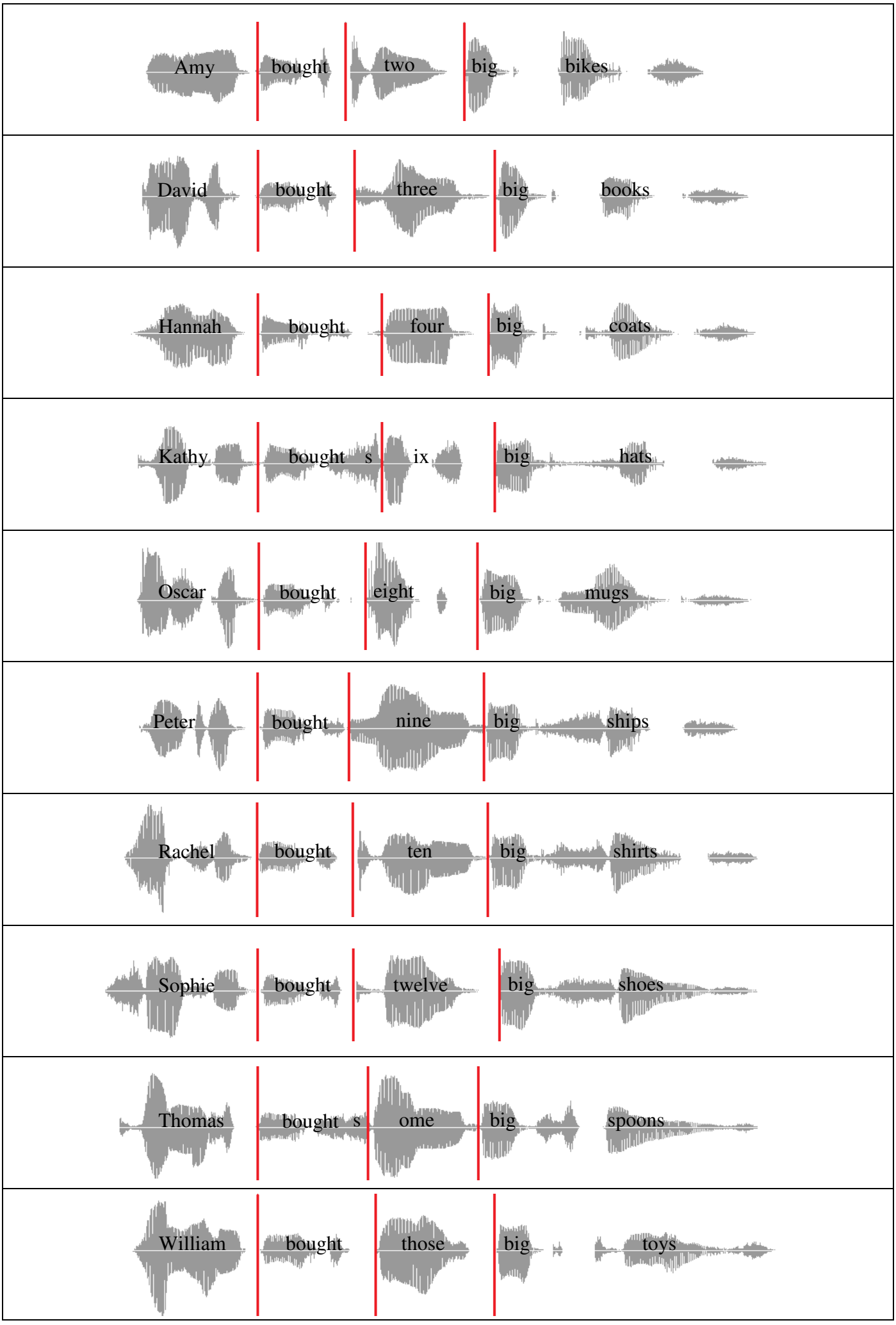

Hod
Hod
Hod
5000
4000
3000
2000
1000
0
Formant frequency (Hz)
F2=1068Hz
F2=1042Hz
F2=1044Hz
F2 Average=1051Hz
F1=652Hz
F1=660Hz
F1=639Hz
F1 Average=650Hz
0
1.37
Time (s)

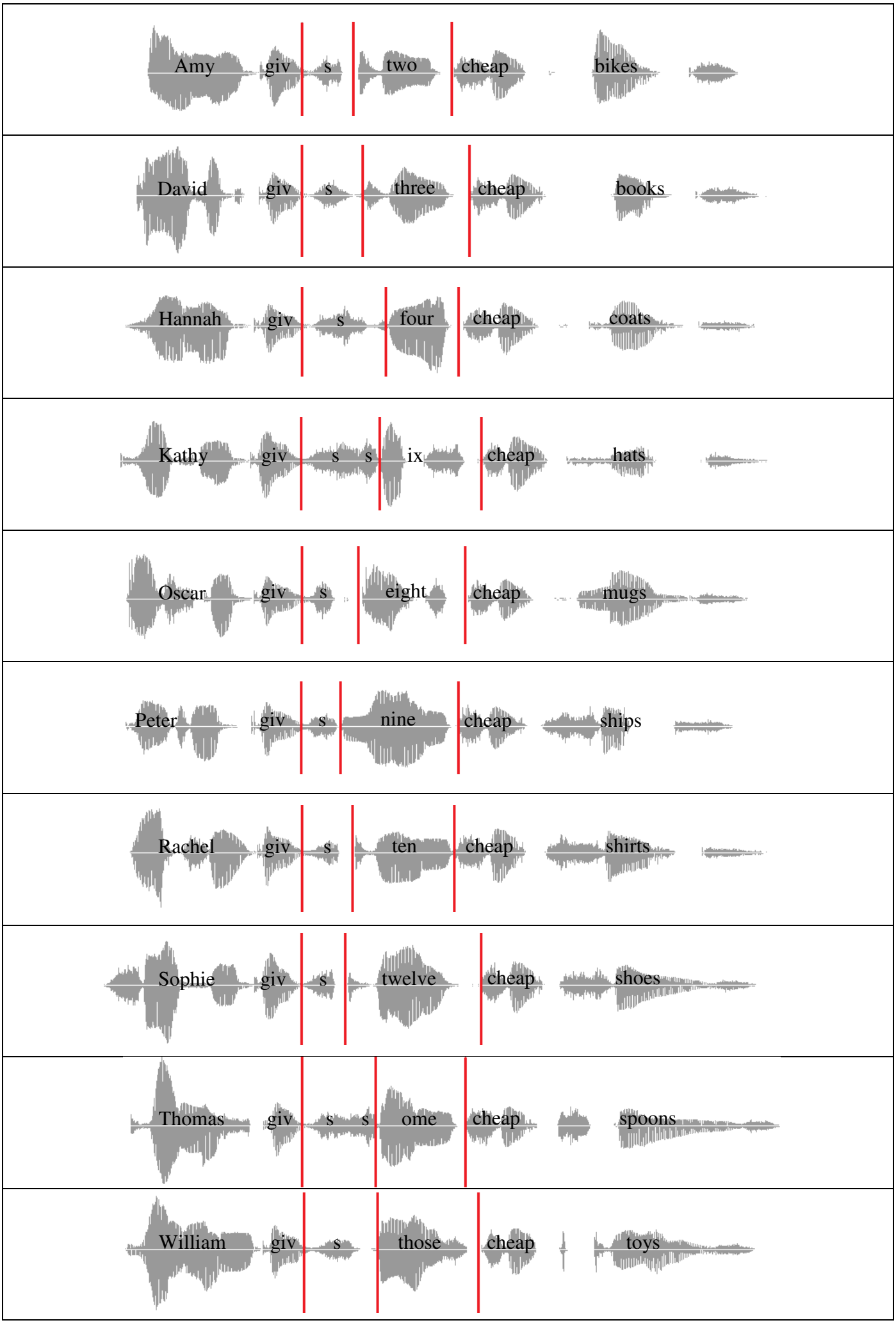

Hood
Hood
Hood
5000
4000
3000
2000
1000
0
Formant frequency (Hz)
F2=1104Hz
F2=1271Hz
F2=1101Hz
F2 Average=1159Hz
F1=549Hz
F1=550Hz
F1=527Hz
F1 Average=542Hz
0
1.123
Time (s)

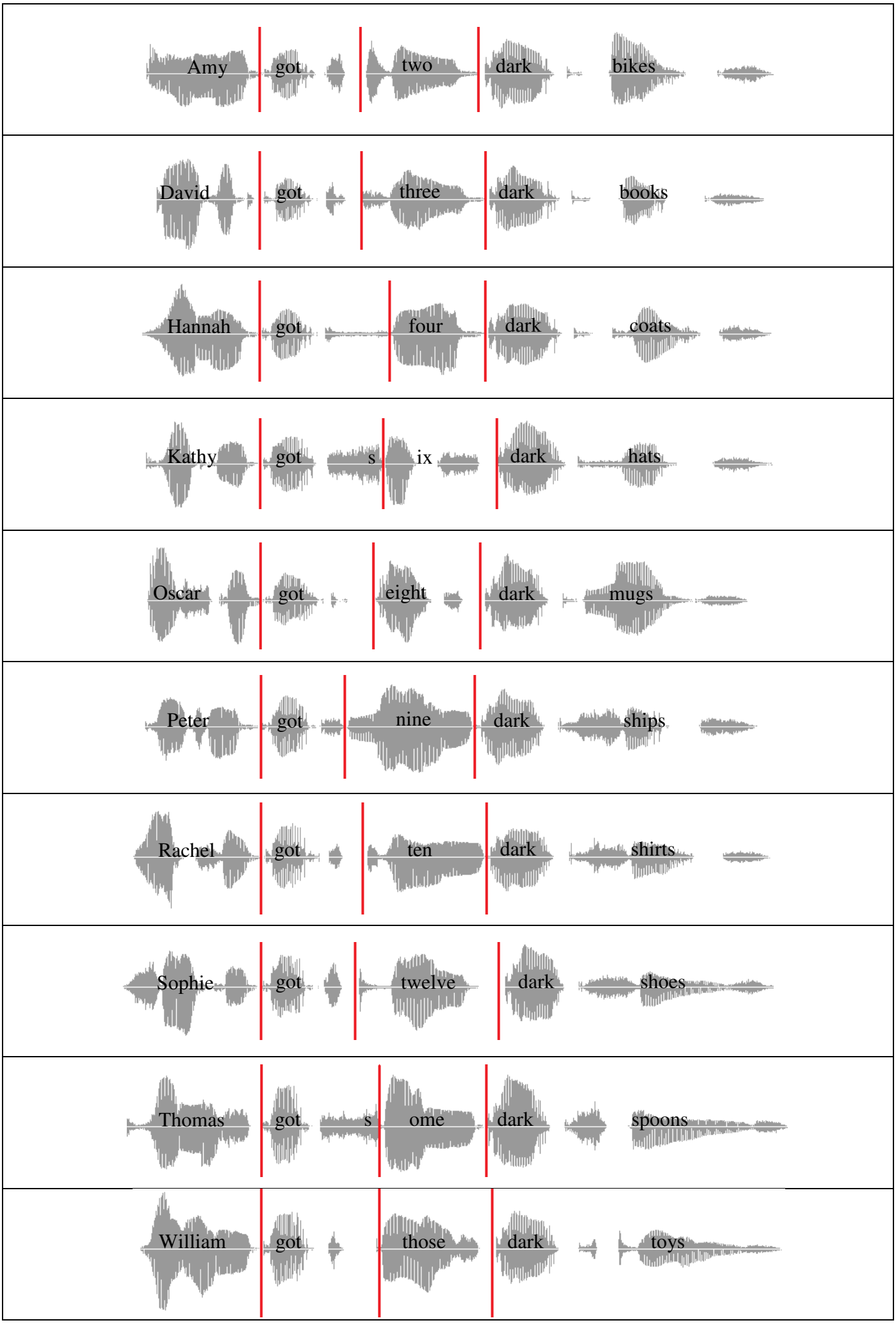

Hud
Hud
Hud
5000
4000
3000
2000
1000
0
Formant frequency (Hz)
F2=1524Hz
F2=1442Hz
F2=1638Hz
F2 Average=1535Hz
F1=841Hz
F1=821Hz
F1=872Hz
F1 Average=845Hz
0
1.461
Time (s)

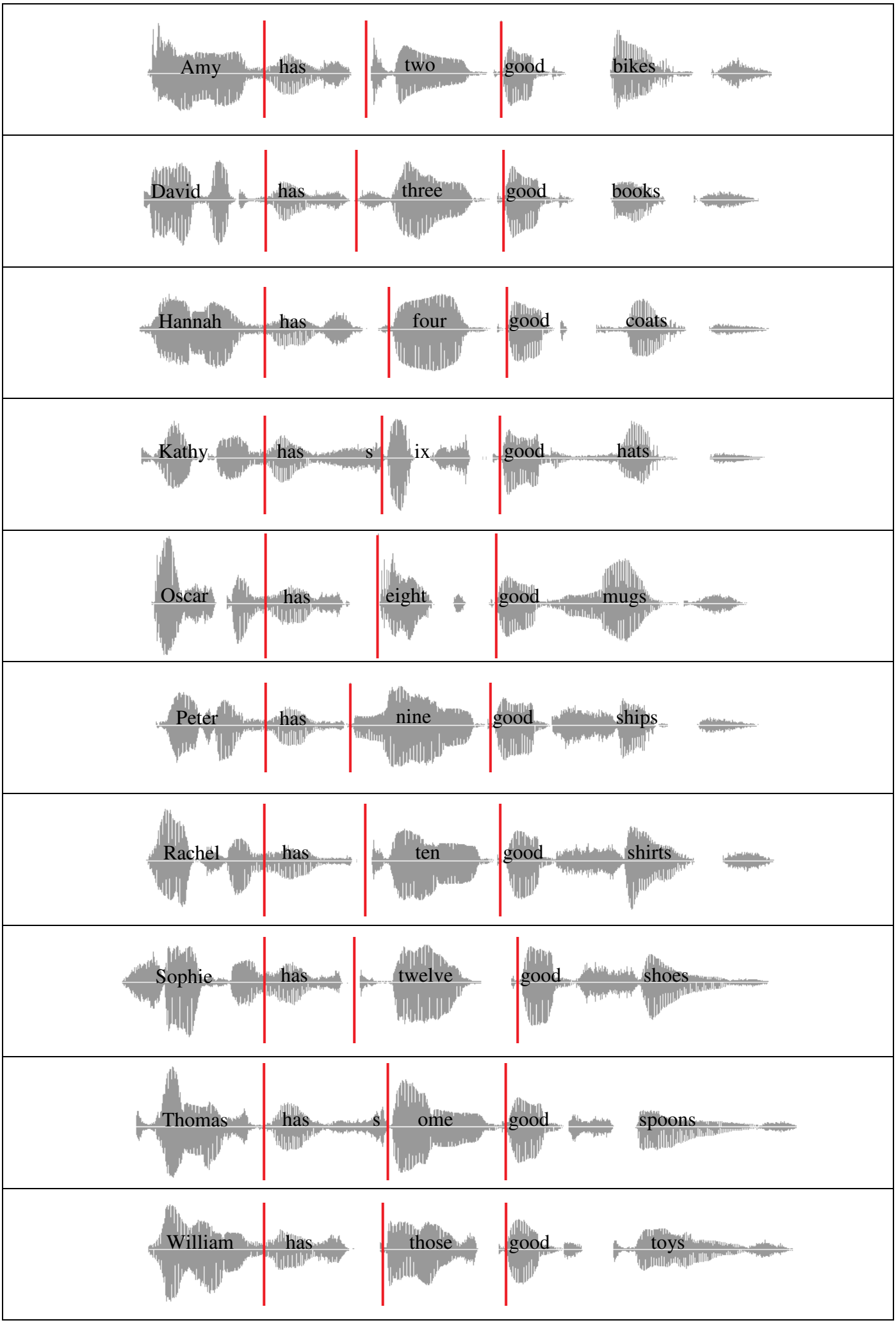

Who'd
Who'd
Who'd
5000
4000
3000
2000
1000
0
Formant frequency (Hz)
F2=1690Hz
F2=1644Hz
F2=1647Hz
F2 Average=1660Hz
F1=438Hz
F1=426Hz
F1=460Hz
F1 Average=441Hz
0
1.54
Time (s)

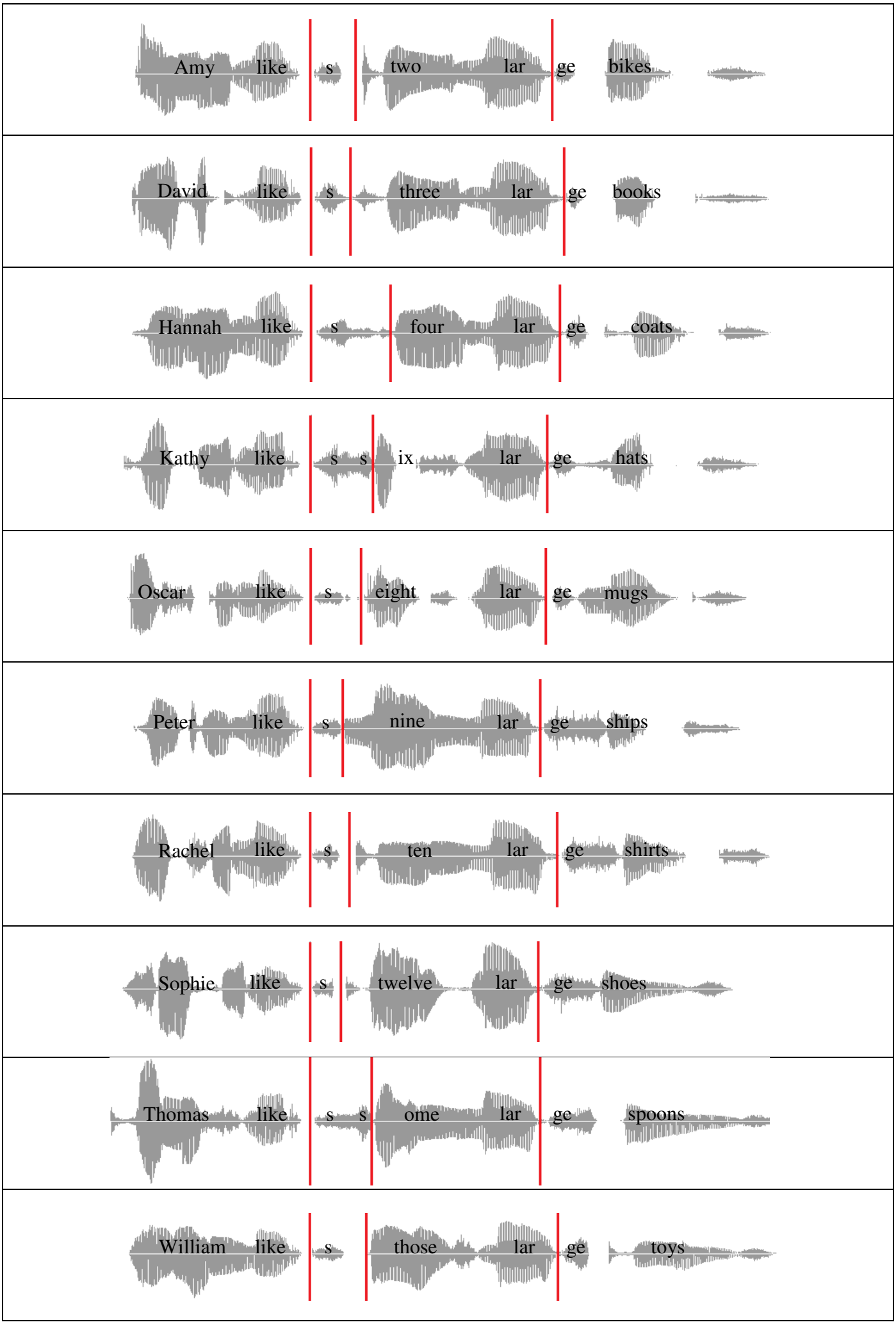
Appendix 4 – Audio-visual Segmentation Points

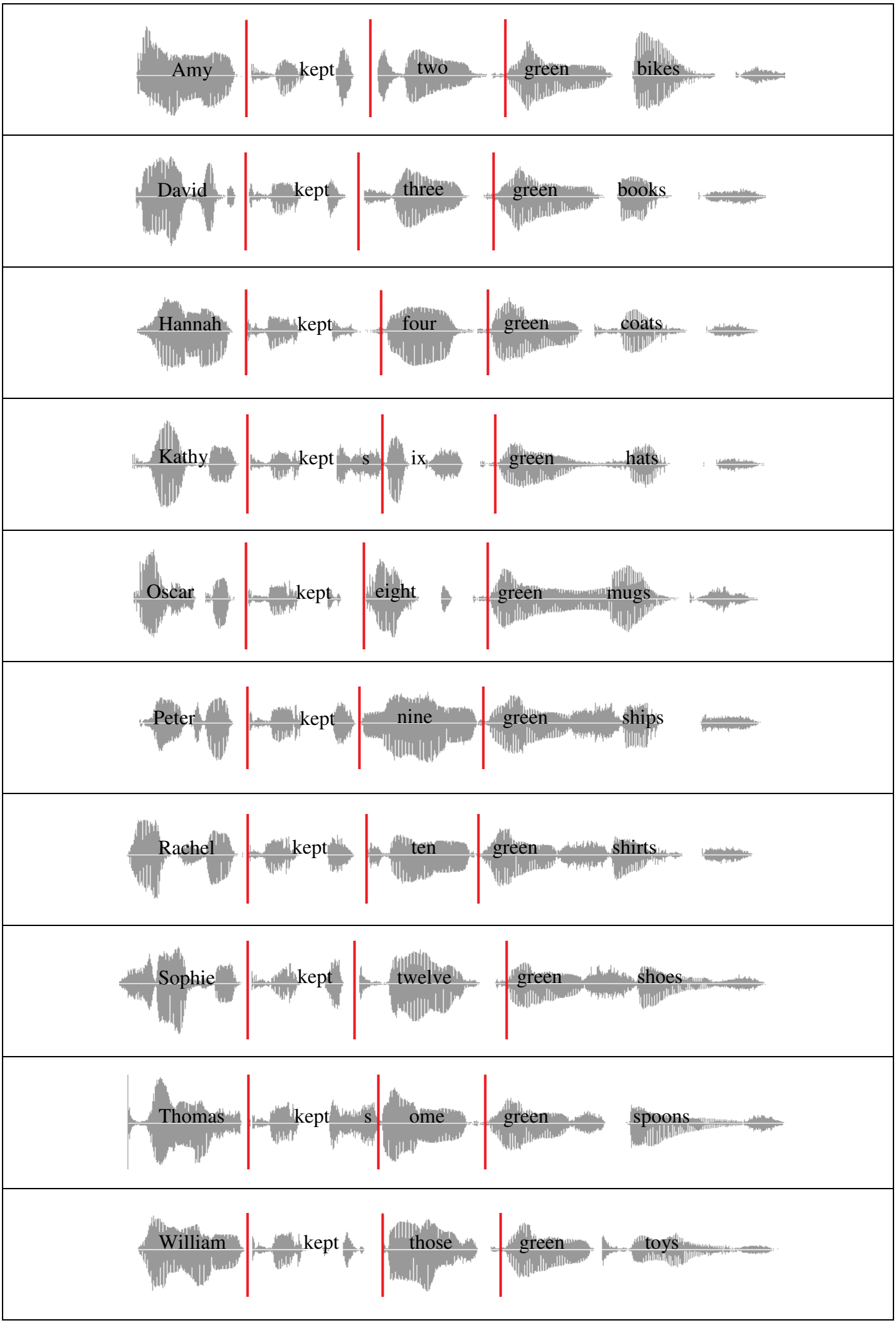

Amy bought two big bikes
David bought three big books
Hannah bought four big coats
Kathy bought s ix big hats
Oscar bought eight big mugs
Peter bought nine big ships
Rachel bought ten big shirts
Sophie bought twelve big shoes
Thomas bought s ome big spoons
William bought those big toys

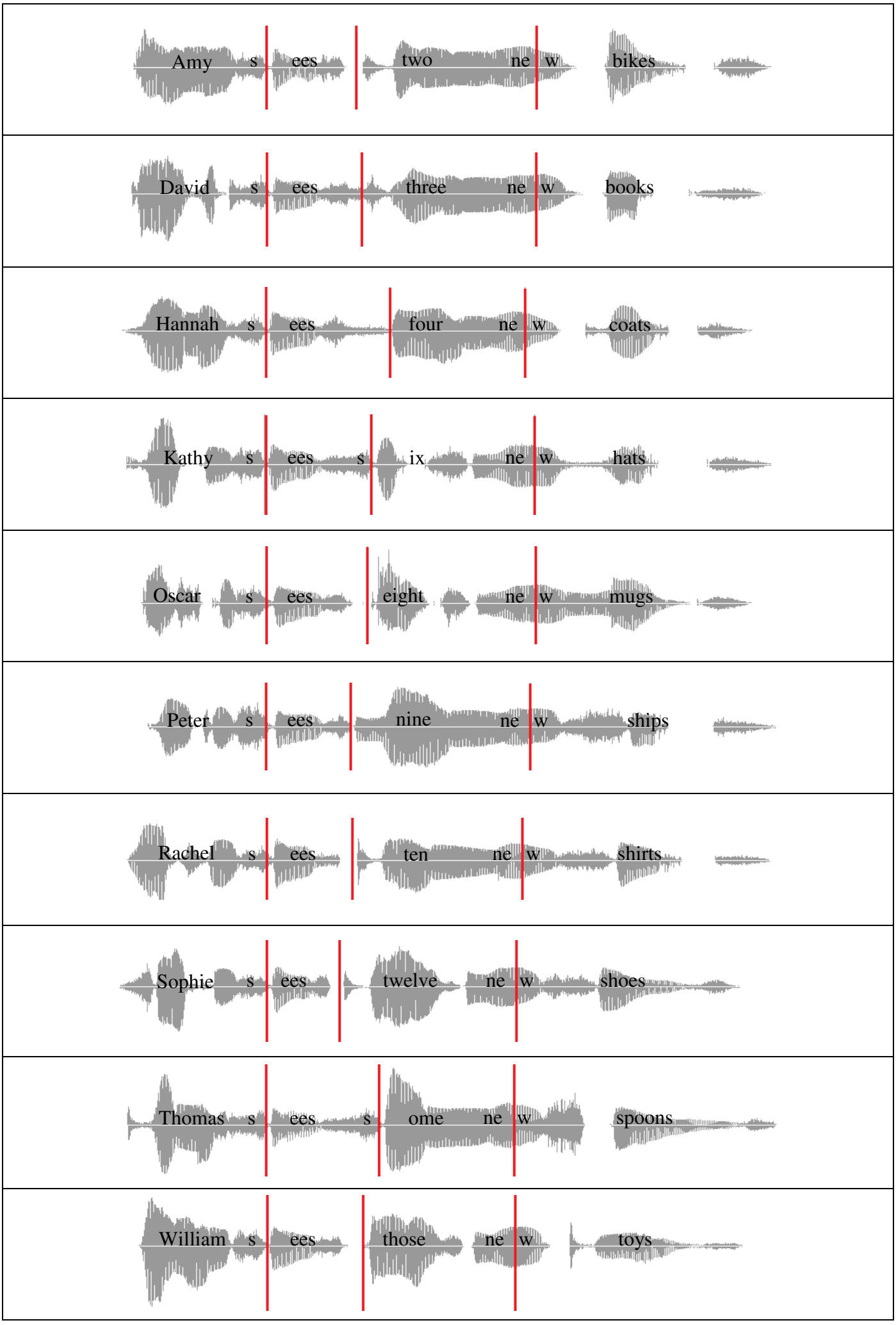

Amy giv s two cheap bikes
David giv s three cheap books
Hannah giv s four cheap coats
Kathy giv s s ix cheap hats
Oscar giv s eight cheap mugs
Peter giv s nine cheap ships
Rachel giv s ten cheap shirts
Sophie giv s twelve cheap shoes
Thomas giv s s ome cheap spoons
William giv s those cheap toys

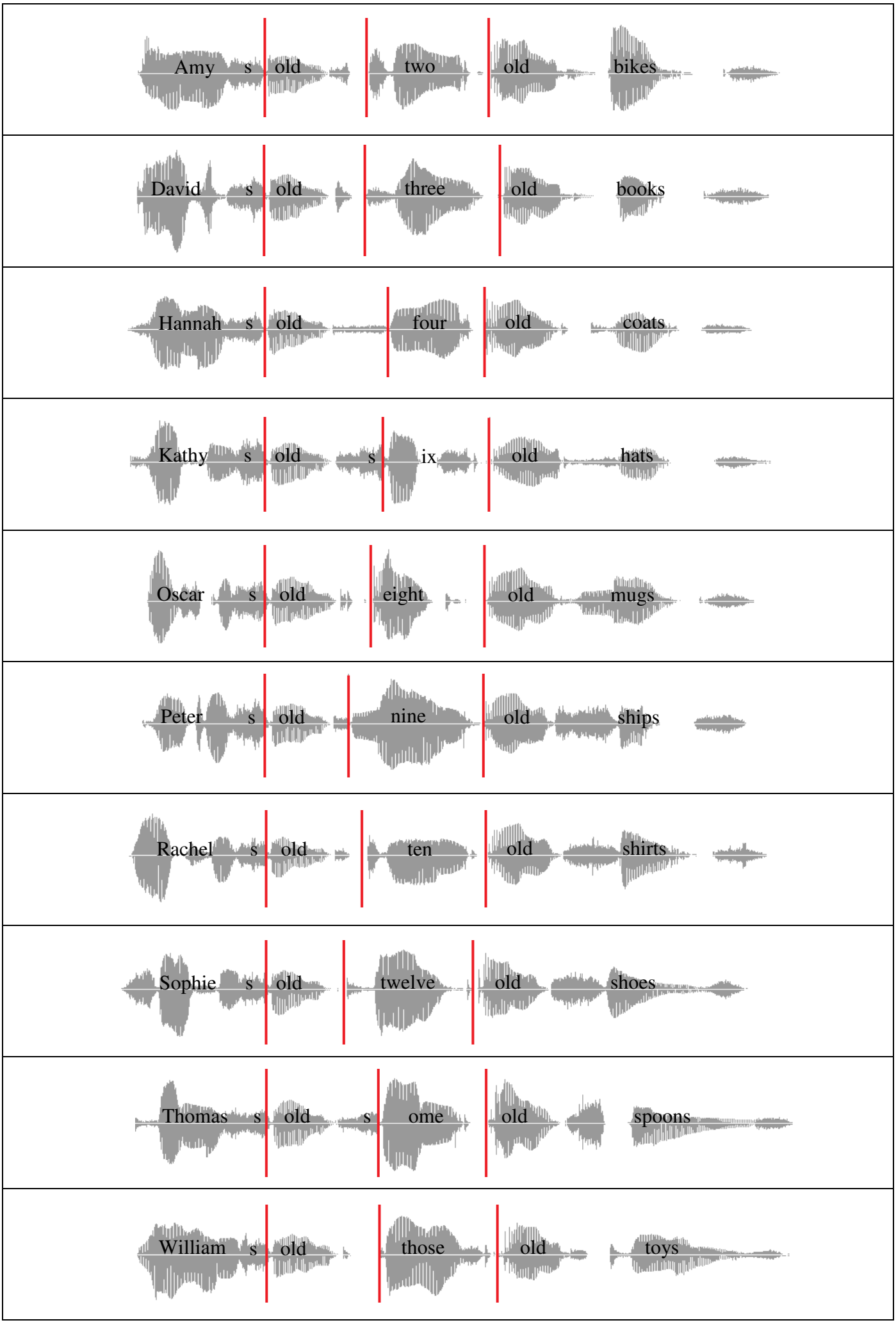

Amy got two dark bikes
David got three dark books
Hannah got four dark coats
Kathy got s ix dark hats
Oscar got eight dark mugs
Peter got nine dark ships
Rachel got ten dark shirts
Sophie got twelve dark shoes
Thomas got s ome dark spoons
William got those dark toys

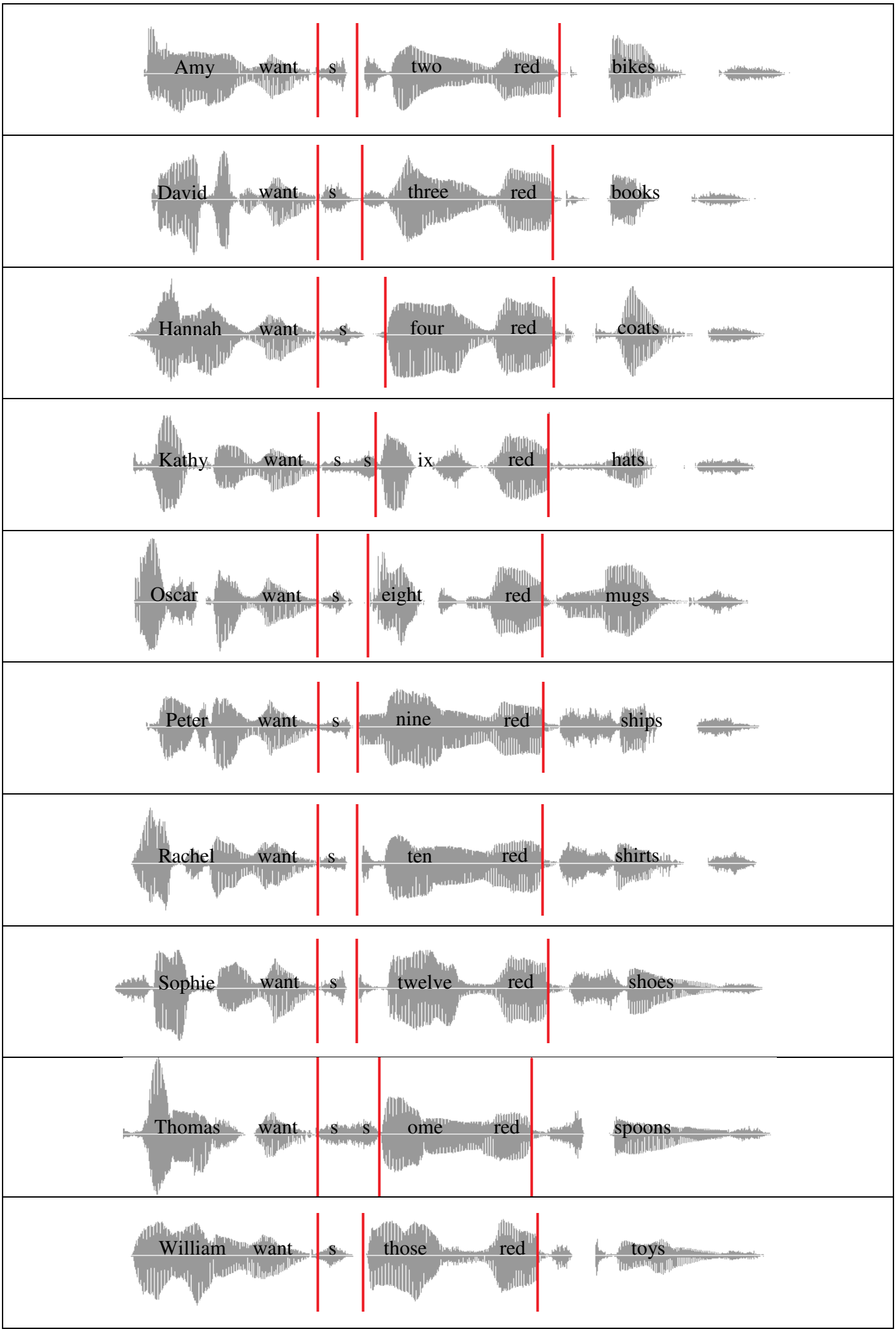

Amy has two good bikes
David has three good books
Hannah has four good coats
Kathy has s ix good hats
Oscar has eight good mugs
Peter has nine good ships
Rachel has ten good shirts
Sophie has twelve good shoes
Thomas has s ome good spoons
William has those good toys

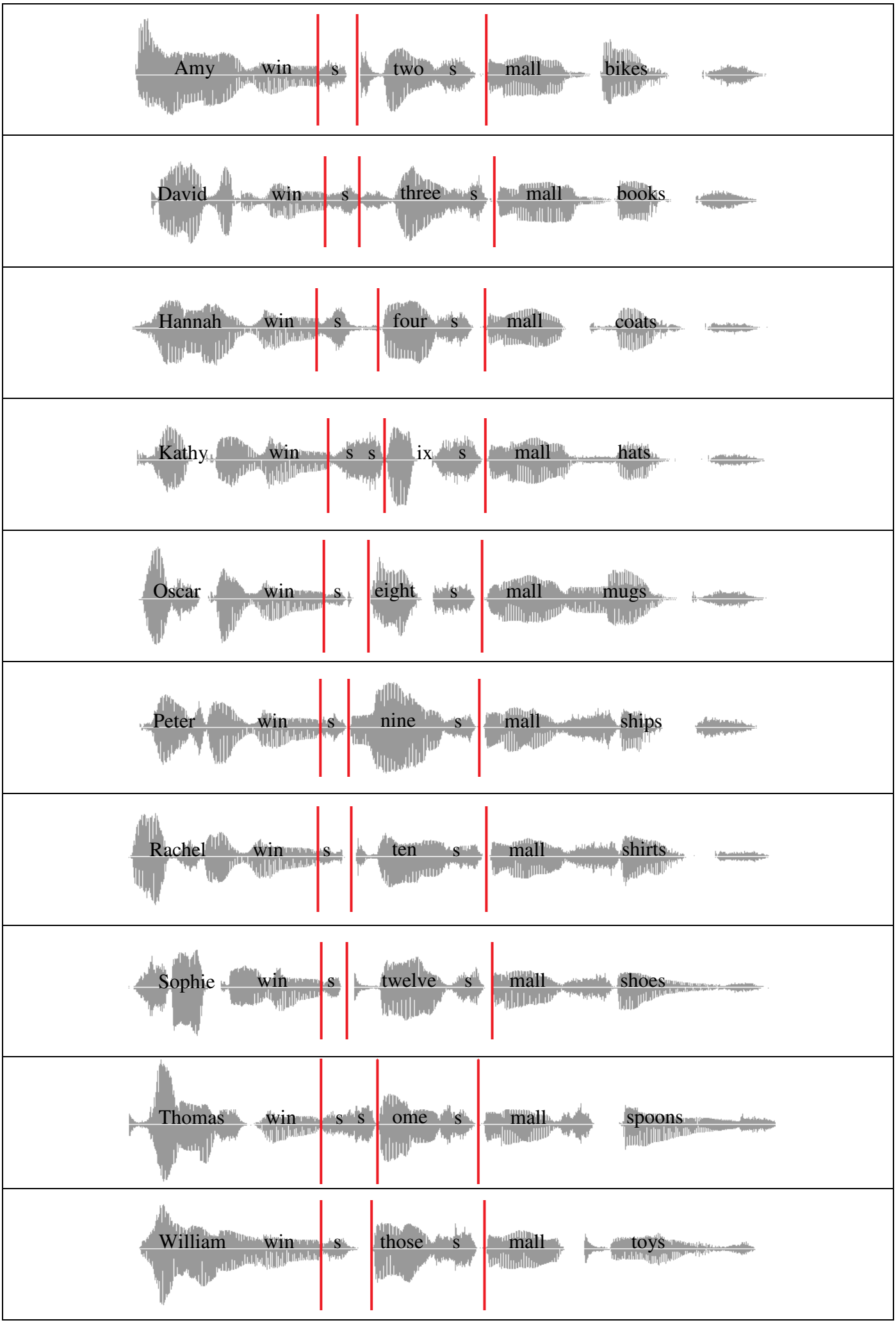

Amy like s two lar ge bikes
David like s three lar ge books
Hannah like s four lar ge coats
Kathy like s s ix lar ge hats
Oscar like s eight lar ge mugs
Peter like s nine lar ge ships
Rachel like s ten lar ge shirts
Sophie like s twelve lar ge shoes
Thomas like s s ome lar ge spoons
William like s those lar ge toys


Amy kept two green bikes
David kept three green books
Hannah kept four green coats
Kathy kept s ix green hats
Oscar kept eight green mugs
Peter kept nine green ships
Rachel kept ten green shirts
Sophie kept twelve green shoes
Thomas kept s ome green spoons
William kept those green toys


Amy s ees two ne w bikes
David s ees three ne w books
Hannah s ees four ne w coats
Kathy s ees s ix ne w hats
Oscar s ees eight ne w mugs
Peter s ees nine ne w ships
Rachel s ees ten ne w shirts
Sophie s ees twelve ne w shoes
Thomas s ees s ome ne w spoons
William s ees those ne w toys


Amy s old two old bikes
David s old three old books
Hannah s old four old coats
Kathy s old s ix old hats
Oscar s old eight old mugs
Peter s old nine old ships
Rachel s old ten old shirts
Sophie s old twelve old shoes
Thomas s old s ome old spoons
William s old those old toys


Amy want s two red bikes
David want s three red books
Hannah want s four red coats
Kathy want s s ix red hats
Oscar want s eight red mugs
Peter want s nine red ships
Rachel want s ten red shirts
Sophie want s twelve red shoes
Thomas want s s ome red spoons
William want s those red toys


Amy win s two s mall bikes
David win s three s mall books
Hannah win s four s mall coats
Kathy win s s ix s mall hats
Oscar win s eight s mall mugs
Peter win s nine s mall ships
Rachel win s ten s mall shirts
Sophie win s twelve s mall shoes
Thomas win s s ome s mall spoons
William win s those s mall toys

Appendix 5 – FFmpeg Command Syntax

Definitions:

-ab 192k = set the audio bit rate to 192 kbits/sec
-an = no audio
-b 1000k = set the maximum total bit rate to 1000 kbits/sec
-i = in file
-s 640x480 = set the picture size to 640 horizontal x 480 vertical pixels
-acodec copy = copy existing audio codec
–newaudio = create a new audio file
-newvideo = create a new video file
-sameq = conversion from mpeg4 to mpg of the same video quality
-vcodec copy = copy existing video codec
-vcodec msmpeg4 = encode video in Microsoft mpeg4 file format
-vn = no video
-y = overwrite without prompting
amy_bought.avi = word pair video file in mpeg4 format
amy_bought_1.avi = uncompressed word pair video file
amy_bought.mpg = word pair video file in mpg format
folder 1 = location of FFmpeg executable
folder 2 = folder containing video files
folder 3 = folder containing audio files

Conversion from uncompressed video to mpeg4 format (in avi container):

"C:\folder 1\ffmpeg.exe" -i "C:\folder 2\amy_bought_1.avi" -y -an -vcodec msmpeg4 -ab 192k -b 1000k -s 640x480 "C:\folder 2\amy_bought.avi"

Conversion from mpeg4 (in avi container) to concatenateable mpg format:

"C:\folder 1\ffmpeg.exe" -i "C:\folder 2\amy_bought.avi" -y -sameq "C:\folder 2\amy_bought.mpg"

Conversion from mpg to mpeg4 format (in avi container)

"C:\folder 1\ffmpeg.exe" -i "C:\folder 2\Visual alone.mpg" -y -an -vcodec msmpeg4 -ab 192k -b 1000k -s 640x480 "C:\folder 2Visual alone.avi"

Combination of audio and video (in avi container)

"C:\folder 1\ffmpeg.exe" -i "C:\folder 2\Visual alone.avi" -i "C:\folder 3\Auditory alone.wav" -y -vcodec copy -an -vn -acodec copy "C:\folder 2Auditory-visual.avi" -newvideo –newaudio

Appendix 6 – MS-DOS Command Syntax

Definitions:

cmd = call MS-DOS command
/b = indicates a binary file
/c = run command and then terminate
amy_bought.mpg = first word pair
bought_two.mpg = second word pair
two_big.mpg = third word pair
big_bikes.mpg = fourth word pair
Visual alone.mpg = concatenated sentence video file

Concatenation of mpg video format:

cmd /c copy /b "amy_bought.mpg"+"bought_two.mpg"+"two_big.mpg"+" big_bikes.mpg" "Visual alone.mpg"

References

Arnold, J. F., Frater, M. R., & Pickering, M. R. (2007). *Digital Television : Technology and Standards*. New Jersey: John Wiley & Sons, Inc.

Beattie, R. C. (1989). Word recognition functions for the CID W-22 test in multitalker noise for normally hearing and hearing-impaired subjects. *The Journal of Speech and Hearing Disorders, 54*(1), 20-32.

Beattie, R. C., Barr, T., & Roup, C. (1997). Normal and hearing-impaired word recognition scores for monosyllabic words in quiet and noise. *British Journal of Audiology, 31*(3), 153-164.

Boothroyd, A., & Nittrouer, S. (1988). Mathematical treatment of context effects in phoneme and word recognition. *The Journal of the Acoustical Society of America, 84*(1), 101-114.

Buss, E., Hall, J. W., Grose, J. H., & Dev, M. B. (2001). A comparison of threshold estimation methods in children 6-11 years of age. *The Journal of the Acoustical Society of America, 109*(2), 727-731.

Carhart, R. (1951). Basic principles of speech audiometry. *Acta Oto-laryngologica, 40*(1-2), 62-71.

Carhart, R., & Tillman, T. W. (1970). Interaction of competing speech signals with hearing losses. *Archives of Otolaryngology, 91*(3), 273-279.

Carhart, R., Tillman, T. W., & Johnson, K. R. (1966). Binaural masking of speech by periodically modulated noise. *Journal of the Acoustical Society of America, 39*, 1037–1050.

Cheng, Y., Liu, Q., Zhu, X., Zhao, C., & Li, S. (2011). Research on Digital Content Protection Technology for Video and Audio Based on FFmpeg. *International Journal of Advancements in Computing Technology, 3*(8), 9-17.

Cox, R. M., Alexander, G. C., & Gilmore, C. (1987). Development of the Connected Speech Test (CST). *Ear and Hearing, 8*(5 Suppl), 119S-126S.

Cropper, S. J., & Derrington, A. M. (1996). Rapid colour-specific detection of motion in human vision. *Nature, 379*(6560), 72-74.

Dirks, D. D., & Bower, D. (1970). Effect of forward and backward masking on speech intelligibility. *Journal of the Acoustic Society of America, 47*, 1003-1008.

Dirks, D. D., Morgan, D. E., & Dubno, J. R. (1982). A procedure for quantifying the effects of noise on speech recognition. *The Journal of Speech and Hearing Disorders, 47*(2), 114-123.

Etymotic Research. (2005). Bamford-Kowal-Bench Speech-in-Noise Test (Version 1.03) [Audio CD]: Elk Grove Village, IL: Author.

Finney, D. J. (1952). *Statistical Methods in Biological Assay*. London: Griffin.

Gifford, R. H., Shallop, J. K., & Peterson, A. M. (2008). Speech recognition materials and ceiling effects: considerations for cochlear implant programs. *Audiology and Neuro-Otology, 13*(3), 193-205.

Grant, K. W., Walden, B. E., & Seitz, P. F. (1998). Auditory-visual speech recognition by hearing-impaired subjects: consonant recognition, sentence

recognition, and auditory-visual integration. *The Journal of the Acoustical Society of America, 103*(5 Pt 1), 2677-2690.

Hagerman, B. (1982). Sentences for testing speech intelligibility in noise. *Scandinavian Audiology, 11*(2), 79-87.

Hall, S. J. (2006). *The Development of a New English Sentence in Noise Test and an English Number Recognition Test.* MSc, University of Southampton.

Hallgren, M., Larsby, B., & Arlinger, S. (2006). A Swedish version of the Hearing In Noise Test (HINT) for measurement of speech recognition. *International Journal of Audiology, 45*(4), 227-237.

Hewitt, D. R. (2007). *Evaluation Of An English Speech-In-Noise Audiometry Test.* MSc, University of Southampton.

Hirsh, I. J., Davis, H., Silverman, S. R., Reynolds, E. G., Eldert, E., & Benson, R. W. (1952). Development of materials for speech audiometry. *The Journal of Speech and Hearing Disorders, 17*(3), 321-337.

Hochmuth, S., Brand, T., Zokoll, M. A., Castro, F. Z., Wardenga, N., & Kollmeier, B. (2012). A Spanish matrix sentence test for assessing speech reception thresholds in noise. *International Journal of Audiology, 51*(7), 536-544.

Hope, R. V. (2010). *Towards the Development of the New Zealand Hearing in Noise Test (NZHINT).* MAud, University of Canterbury, Christchurch.

Howard-Jones, P. A., & Rosen, S. (1993). The perception of speech in fluctuating noise. *Acustica, 78*, 258-272.

Killion, M. C., Niquette, P. A., Gudmundsen, G. I., Revit, L. J., & Banerjee, S. (2004). Development of a quick speech-in-noise test for measuring signal-to-noise ratio loss in normal-hearing and hearing-impaired listeners. *The Journal of the Acoustical Society of America, 116*(4 Pt 1), 2395-2405.

King, S. M. (2011). *Development and Evaluation of a New Zealand Digit Triplet Test for Auditory Screening.* MAud, University of Canterbury, Christchurch.

Kollmeier, B., & Wesselkamp, M. (1997). Development and evaluation of a German sentence test for objective and subjective speech intelligibility assessment. *The Journal of the Acoustical Society of America, 102*(4), 2412-2421.

Leek, M. R. (2001). Adaptive procedures in psychophysical research. *Perception and Psychophysics, 63*(8), 1279-1292.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *The Journal of the Acoustical Society of America, 49*(2), 467-477.

Levitt, H. (1978). Adaptive testing in audiology. *Scandinavian Audiology*(6), 241-291.

Luts, H., Boon, E., Wable, J., & Wouters, J. (2008). FIST: a French sentence test for speech intelligibility in noise. *International Journal of Audiology, 47*(6), 373-374.

Maclagan, M., & Hay, J. (2007). Getting fed up with our feet: Contrast maintenance and the New Zealand English front vowel shift. *Language Variation and Change, 19*(01), 1-25.

McArdle, R., & Hnath-Chislom, T. (2009). Speech Audiometry. In J. Katz (Ed.), *Handbook of Clinical Audiology* (6th ed., pp. 64-79). Baltimore: Lippincott Williams & Wilkins.

McArdle, R. A., Wilson, R. H., & Burks, C. A. (2005). Speech recognition in multitalker babble using digits, words, and sentences. *Journal of the American Academy of Audiology, 16*(9), 726-739.

Mendel, L. L. (2008). Current considerations in pediatric speech audiometry. *International Journal of Audiology, 47*(9), 546-553.

Miller, G. A. (1947). The masking of speech. *Psychological Bulletin, 44*, 105-129.

Miller, G. A., & Licklider, J. C. R. (1950). The intelligibility of interrupted speech. *Journal of the Acoustic Society of America, 22*, 167-173.

Mitchell, A. G. (1946). *The Pronunciation of English in Australia*. Sydney: Angus & Robertson.

Nilsson, M., Soli, S. D., & Sullivan, J. A. (1994). Development of the Hearing in Noise Test for the measurement of speech reception thresholds in quiet and in noise. *The Journal of the Acoustical Society of America, 95*(2), 1085-1099.

Niquette, P., Arcaroli, J., Revit, L., Parkinson, A., Staller, S., Skinner, M. (2003). *Development of the BKB-SIN Test*. Paper presented at the annual meeting of the American Auditory Society, Scottsdale, AZ.

O'Beirne, G. A., McGaffin, A. J., & Rickard, N. A. (2012). Development of an adaptive low-pass filtered speech test for the identification of auditory processing disorders. *International Journal of Pediatric Otorhinolaryngology, 76*(6), 777-782.

Orchik, D. J., Krygier, K. M., & Cutts, B. P. (1979). A comparison of the NU-6 and W-22 speech discrimination tests for assessing sensorineural hearing loss. *The Journal of Speech and Hearing Disorders, 44*(4), 522-527.

Ozimek, E., Kutzner, D., Sek, A., & Wicher, A. (2009). Polish sentence tests for measuring the intelligibility of speech in interfering noise. *International Journal of Audiology, 48*(7), 433-443.

Ozimek, E., Warzybok, A., & Kutzner, D. (2010). Polish sentence matrix test for speech intelligibility measurement in noise. *International Journal of Audiology, 49*(6), 444-454.

Plomp, R., & Mimpen, A. M. (1979). Improving the reliability of testing the speech reception threshold for sentences. *Audiology, 18*(1), 43-52.

Pollack, I. (1954). Masking of speech by repeated bursts of noise. *Journal of the Acoustical Society of America, 26*, 1053–1055.

Pollack, I. (1955). Masking by periodically interrupted noise. *Journal of the Acoustical Society of America, 27*, 353–355.

Porter, T., & Duff, T. (1984). Compositing digital images. *Computer Graphics, 18*(3), 253-259.

Rowland, J. P., Dirks, D. D., Dubno, J. R., & Bell, T. S. (1985). Comparison of speech recognition-in-noise and subjective communication assessment. *Ear and Hearing, 6*(6), 291-296.

Smits, C., Kapteyn, T. S., & Houtgast, T. (2004). Development and validation of an automatic speech-in-noise screening test by telephone. *International Journal of Audiology, 43*(1), 15-28.

Sommers, M. S., Tye-Murray, N., & Spehar, B. (2005). Auditory-visual speech perception and auditory-visual enhancement in normal-hearing younger and older adults. *Ear and Hearing, 26*(3), 263-275.

Strom, K. E. (2006). The HR 2006 dispenser survey. *The Hearing Review, 13*(6), 16–39.

Stuart, A., & Phillips, D. P. (1996). Word recognition in continuous and interrupted broadband noise by young normal-hearing, older normal-hearing, and presbycusic listeners. *Ear and Hearing Research, 17*, 478–489.

Stuart, A., & Phillips, D. P. (1998). Deficits in auditory temporal resolution revealed by a comparison of word recognition under interrupted and continuous noise masking. *Seminars in Hearing, 19*, 333–344.

Sumby, W. H., & Pollack, I. (1954). Visual contributions to speech intelligibility in noise. *Journal of the Acoustical Society of America, 26*, 212-215.

Swosho. (2012). Head alignment clamp Retrieved 18 August, 2012, from http://www.thingiverse.com/thing:22852

Tye-Murray, N., Sommers, M., & Spehar, B. (2007). Auditory and visual lexical neighborhoods in audiovisual speech perception. *Trends in Amplification, 11*(4), 233-241.

Tye-Murray, N., Sommers, M., Spehar, B., Myerson, J., Hale, S., & Rose, N. S. (2008). Auditory-visual discourse comprehension by older and young adults in favorable and unfavorable conditions. *International Journal of Audiology, 47*(s2), S31-S37.

Versfeld, N. J., Daalder, L., Festen, J. M., & Houtgast, T. (2000). Method for the selection of sentence materials for efficient measurement of the speech reception threshold. *The Journal of the Acoustical Society of America, 107*(3), 1671-1684.

Wagener, K., Brand, T., & Kollmeier, B. (1999b). [Development and evaluation of a German sentence test II: Optimization of the Oldenburg sentence test]. *Zeitschrift für Audiologie, 38*, 44-56.

Wagener, K., Brand, T., & Kollmeier, B. (1999c). [Development and evaluation of a German sentence test III: Evaluation of the Oldenburg sentence test]. *Zeitschrift für Audiologie, 38*, 86-95.

Wagener, K., Josvassen, J. L., & Ardenkjaer, R. (2003). Design, optimization and evaluation of a Danish sentence test in noise. *International Journal of Audiology, 42*(1), 10-17.

Wagener, K., Kühnel, V., & Kollmeier, B. (1999a). [Development and evaluation of a German sentence test I: Design of the Oldenburg sentence test]. *Zeitschrift für Audiologie, 38*, 4-15.

Wells, J. C. (1982). *Accents of English* (Vol. 1). Cambridge: Cambridge University Press.

Wilson, R. H., & Carhart, R. (1969). Influence of pulsed masking on the threshold for spondees. *The Journal of the Acoustical Society of America, 46*(4), 998-1010.

Wilson, R. H., McArdle, R. A., & Smith, S. L. (2007). An Evaluation of the BKB-SIN, HINT, QuickSIN, and WIN Materials on Listeners With Normal Hearing and Listeners With Hearing Loss. *Journal of Speech, Language, and Hearing Research, 50*(4), 844-856.

Zokoll, M. A., Wagener, K. C., Brand, T., Buschermöhle, M., & Kollmeier, B. (2012). Internationally comparable screening tests for listening in noise in several European languages: The German digit triplet test as an optimization prototype. *International Journal of Audiology, 51*(9), 697-707.